1996

Detection And Estimation Of Boundaries In Spatial Data

Lei Xie

DETECTION AND ESTIMATION OF BOUNDARIES IN SPATIAL DATA

by

Lei Xie

Department of Statistical and Actuarial Sciences

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario
May 1996

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

ISBN 0-612-15091-7

Canada

ABSTRACT

The objective of this thesis is to develop methodology for detecting and estimating parameter changes at unknown boundaries for regression models of spatial data. This methodology has many applications including quality control, epidemiology, pharmacology, agriculture, meteorology and geology.

Statistics are derived for testing a spatial array of observations from a regression model for no change in parameters against possible alternatives involving change of parameters at unknown boundaries based on a Bayes-type approach and a locally optimal framework. These test statistics are defined in terms of set indexed partial sums of regression residuals. Limit processes are obtained for set indexed partial sums of regression residuals with i.i.d. errors, and distributional results based on these processes are obtained for selected change-boundary statistics. Limit processes are derived for a matrix array of partial sums of regression residuals with stationary spatial error structure. The relationship between these procedures and those for i.i.d. errors is established. Methods are given for estimating the locations of boundaries separating spatial sets of observations which are characterized by models with different parameter sets.

This methodology is then applied to the Mercer and Hall wheat-yield data and to age-period breast cancer mortality data.

ACKNOWLEDGEMENTS

I would like to acknowledge my indebtedness to Professor Ian B. MacNeill, my supervisor, for his outstanding guidance and constant encouragement throughout all stages of this research project. His helpful advice, valuable comments and patient discussions have considerably improved the present thesis. I feel greatly privileged to have been able to work under his supervision and I can never express my gratitude in words to him.

I am extremely grateful to Dr. Gary Umphrey for his help and encouragement during my studies at Western.

I am also extremely thankful to Dr. M.S. Haq and Dr. S.B. Provost for their suggestions and comments.

Finally, I would also like to thank my parents, my wife and my daughter for their patience and support.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

Statistical models for time series data are generally characterized by several unknown parameters. These parameters may change over time, and if the changes occur unannounced and at unknown time points, then the associated problems of detection and estimation are referred to as the change-point problem. This problem has been extensively investigated by several authors during the last few decades.

Recently, change-point methods have been extended to spatial data. Carlstein and Krishnamoorthy (1992) have considered the problem of estimating the location of an unknown boundary given that a boundary is present. They use a non-parametric approach that uses empirical distribution functions for observations assumed to be from one distribution on one side of the boundary and from a different distribution on the other side. MacNeill and Jandhyala (1993) extended some of the parametric change-point methods developed for univariate time series to the problem of testing spatial data for parameter changes at unknown location, and discussed the problem

of estimating the location of boundaries between regions described by different sets of parameter values. They considered some aspects of detection and location of boundaries for linear regression.

In this thesis, the problems considered are those of detection and estimation of parameter changes at unknown boundaries in spatial data for regression models with i.i.d. errors and with spatially correlated errors. Several statistics, based on a Bayes-type approach and a locally optimal framework, are derived for testing a spatial array of observations from a regression model for no change in parameters against possible alternatives involving change of parameters at unknown boundaries. These test statistics are defined in term of set indexed partial sums of regression residuals. Limit processes are obtained for set indexed partial sums of regression residuals, and distributional results based on these are obtained for selected change-boundary statistics.

The problem of detection of parameter changes in a sequence of observations at unknown times was first considered by Shewhart (1930) who introduced the control chart approach to quality management in industrial processes. Page (1954, 1955, 1961) formalized this approach by using a sequential test based on cumulative sum (CUSUM) schemes. Chernoff and Zacks (1964) proposed a one-sided Bayes-type test for change of parameter at unknown times in sequences of random variables, and this approach was adapted by Gardner (1969), Sen and Srivastava (1973, 1975) and MacNeill (1974). For regression models, MacNeill (1978a,b) proposed a test based on cusums of residuals, and Jandhyala and MacNeill (1989, 1991) derived Bayes-type

change detection statistics based on partial sums of residuals and discussed their asymptotic distributions for general regression. Tang and MacNeill (1993) provided adjustments to change-point test statistics to account for the effect of serial correlation.

Approaches to the change-point problem based directly on the likelihood function have been proposed by Quandt (1958, 1960), Sen and Srivastava (1975), Hawkins (1977), Worsley (1979, 1983a,b), Hinkley (1970, 1971, 1972), and Esterby and El-Shaarawi (1981). Non-parametric approaches to the problem have been proposed by Bhattacharya and Johnson (1968), Pettit (1979), Bhattacharya and Friesson (1981), Schechtman (1982) and Eastwood (1993). Change detection methods using information criteria have been proposed by Akaike (1974). Bayesian methods were proposed by Smith (1975), Lee and Heghinian (1977), and Hsu (1979). Broemeling and Tsurumi (1987) discussed Bayesian methods for structural change problems.

The problem of detection of parameter changes at unknown boundaries in spatial data from regression models is formulated in Chapter 2. Chernoff and Zacks (1964) first introduced the Bayes-type approach to detecting one-sided changes of parameters and Gardner (1969) extended this method to the two-sided case. In Chapter 2, we derive the Bayes-type statistics for testing one-sided and two-sided changes in parameters of regression models at unknown boundaries, thus extending the results of Jandhyala and MacNeill (1991) to spatial data. The statistics derived by Chernoff and Zacks for testing in the case of a one-sided change of parameter and by Gardner for the two-sided case are special cases for time sequences. We also discuss several

test statistics for some simple boundaries.

Jandhyala and MacNeill (1992) showed that the locally best statistic to test for the constancy of a single parameter of a linear regression discussed by Nabeya and Tanaka (1988) under a random walk alternative and a Bayes-type statistic derived by Jandhyala and MacNeill (1991) under a change-point alternative are identical. In Chapter 3, we apply a locally optimal framework to develop methodology to test a spatial array of observations from a regression model for no change in parameter coefficients against possible alternatives involving change of parameter coefficients at unknown boundaries and we discuss optimality properties of these tests. The test statistics derived by the locally optimal framework in Chapter 3 are the same as those derived by the Bayes-type approach in Chapter 2.

Pyke (1973, 1983), Bass and Pyke (1984) and Alexander and Pyke (1986) considered an array of i.i.d. random variables indexed by the d-dimensional positive integer lattice and developed weak convergence theory for set indexed partial sum processes. Our test statistics are defined in terms of set indexed partial sums of regression residuals. The regression models in spatial data with i.i.d. errors are considered in Chapter 4. We indicate how weak convergence for set indexed partial sum processes can be applied to obtain large sample theory for set indexed partial sums of regression residuals and large sample distributions for our test statistics, and we also obtain the distributional results based on these limit processes for selected change-boundary statistics. Therefore, we extend the results of MacNeill (1978b), Jandhyala and MacNeill (1989), and Tang and MacNeill (1992) to spatial data.

In Chapter 5, we consider the linear regression models with spatially correlated errors. Limit processes for a matrix array of partial sums of residuals from stationary spatial series are first derived, and these results are applied to obtain limit processes for a matrix array of partial sums of regression residuals with stationary spatial error structure. The class of statistics based on these partial sums is used for detection of interventions in spatial series occurring at one of a set of specific unknown boundaries (see Figure 1). The use of these detection statistics for regression with i.i.d. error structures are extended to the case of stationary spatial processes. The results of Tang and MacNeill (1993) are extended to spatial data in Chapter 5.

When the existence of a boundary is identified, the next problem to be considered is estimation of the boundary's location in Chapter 6. Carlstein and Krishnamoorthy (1992) proposed a non-parametric approach to estimating the location of a change-boundary. Methods based on the likelihood approach for estimating change-points are extended to estimation of the locations of boundaries separating spatial sets of observations which are characterized by models with different parameter sets.

In Chapter 7, we discuss the application of this methodology to the Mercer and Hall (1911) wheat-yield data and to Canadian age-period breast cancer mortality data.

Chapter 2

TESTS FOR CHANGES OF PARAMETERS AT UNKNOWN BOUNDARIES FOR REGRESSION MODELS

2.1 Introduction

Chernoff and Zacks (1964) first introduced Bayes-type statistics for one-sided tests of parameter changes at unknown times in the mean of a sequence of independent normal random variables. The Bayes-type approach to deriving test statistics consists first in assuming appropriate prior distributions on nuisance parameters associated with both the null and alternative hypotheses. Then, the respective unconditional likelihood functions are obtained through elimination of the nuisance parameters, and a likelihood ratio statistic is derived. This approach was adapted by Gardner (1969) for two-sided tests of parameter changes in a sequence of normal variables and by MacNeill (1974) for two two-sided tests when the variables are from a one-parameter exponential family. For regression models, MacNeill (1978a,b) proposed a test based on cusums of residuals, and Jandhyala and MacNeill (1989, 1991) derived Bayes-type change detection statistics based on partial sums of residuals. MacNeill and Jandhyala (1993) extended some of the parametric change-point methods developed

for univariate time series to the problem of testing spatial data for parameter changes at unknown boundaries.

This chapter is devoted to the derivation of Bayes-type statistics for the change-boundary problem in regression models for spatial data. In Section 2.2, Bayes-type statistics are derived for testing an array of observations from a regression model for no change in parameters against possible alternatives involving changes at unknown boundaries. We discuss several test statistics for special cases in Section 2.3.

2.2 Tests for Changes of Parameters at Unknown Boundaries in a Regression Model

To begin we consider n^d observations taken on a regular lattice in the unit cube $I^d = [0,1]^d$. The observation at point $n^{-1}\mathbf{j} = (j_1/n, \ldots, j_d/n)$ is denoted by $Y_{\mathbf{j}}$. To define a general regression model, we let $f_k(\cdot, \ldots, \cdot)$, $k = 0, 1, \ldots, p-1$, be a set of d-variate non-stochastic regressor functions, let $\leq$ denote the coordinate-wise partial ordering on set of d-tuples positive integers, and let $\{\varepsilon_{\mathbf{j}} : \mathbf{j} \leq n\mathbf{1}\}$ be an array of iid real-valued random variables from the $N(0,1)$ distribution. Then we define an array of dependent variables $\{Y_{\mathbf{j}} : \mathbf{j} \leq n\mathbf{1}\}$ on a regular lattice on the d-dimensional unit cube I^d as follows:

$$Y_{\mathbf{j}} = \sum_{k=0}^{p-1} \beta_k f_k(n^{-1}\mathbf{j}) + \varepsilon_{\mathbf{j}} ,$$

where $\mathbf{j} = (j_1, \ldots, j_d)$ is a vector of positive integers.

If we denote the vector of regression coefficients by $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$, the design matrix by $\mathbf{X}$, the vector of observations by $\mathbf{Y}$ and the vector of noise variables

by ϵ, then the model may be written in matrix form as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{(stacked)}.$$

The problem of present interest is to test for changes in the parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ at unknown boundaries.

For the purposes of testing we consider a boundary B to be defined by the convex hull of a set of lattice points in the d-dimensional unit cube. This set, plus the lattice points within the convex hull, is denoted by R_B, and the complement of R_B in the unit cube is denoted by $R_{\bar{B}}$. The collection of such boundaries is denoted by $\mathcal{B}$. For large sample distribution theory and for practical application we shall restrict attention to certain subsets of the set of convex subsets of the d-dimensional unit cube.

We wish to test the hypothesis of no change in parameters against certain alternatives involving changes in the parameters at unknown boundary $B \in \mathcal{B}$. To specify alternatives we let δ be the amount of change, and we let $\omega_k(B)$ be 1 or 0 according to whether there is or is not a change at B in parameter β_k.

The problem then is to test if the regression model has changed to

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{H}_B \mathbf{1}_p \delta + \boldsymbol{\epsilon}\ ,$$

where $\mathbf{H}_B = (\omega_0(B) f_0(n^{-1}\mathbf{j}), \ldots, \omega_{p-1}(B) f_{p-1}(n^{-1}\mathbf{j}) : n^{-1}\mathbf{j} \in R_B)$ and $\mathbf{1}_p = (1, \ldots, 1)'$ is a $p \times 1$ vector.

Hence, for the one-sided change case, we wish to test the null hypothesis

$$H_0 : \delta = 0$$

against the alternative hypothesis

$$H_1 : \delta > 0$$

for some unknown boundary B (i.e. $n^{-1}\mathbf{j} \in R_B$).

A likelihood ratio statistic is derived in the sequel for testing these hypotheses using the Bayesian method introduced by Chernoff and Zacks (1964) and which was discussed for regression problems by Jandhyala and MacNeill (1991). The method is adapted to the case of spatial data. This method imposes a priori distributions on the unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\delta} = \mathbf{1}_p\delta$ and boundary B; the choices for this case are as follows:

$$\boldsymbol{\beta} \sim N_p(0, \tau^2\mathbf{I}_p) \text{ and } \delta \sim \frac{1}{2}N(0, \theta^2) .$$

The distributions of $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ and ϵ are assumed to be independent. Also, let $p(\omega_B)$ denote a priori probability distribution of the unknown boundary B, where $\omega_B = (\omega_0(B), \ldots, \omega_{p-1}(B))$.

Under H_0, the likelihood of $\{Y_\mathbf{j} : \mathbf{j} \leq n\mathbf{1}\}$ is given by

$$L_0(\mathbf{Y}) = \left(\frac{1}{2\pi}\right)^{\frac{n^d}{2}} \frac{1}{|\boldsymbol{\Sigma}_0|^{1/2}} \exp\left\{-\frac{1}{2}\mathbf{Y}'\boldsymbol{\Sigma}_0{}^{-1}\mathbf{Y}\right\} ,$$

where $\boldsymbol{\Sigma}_0 = \mathbf{I} + \tau^2\mathbf{X}\mathbf{X}'$.

Note that, under the alternative hypothesis H_1,

$$(\mathbf{Y} \mid \boldsymbol{\beta}, \delta, \omega_B) \sim N_{n^d}(\mathbf{X}\boldsymbol{\beta} + \delta\mathbf{h}_B, \mathbf{I}) ,$$

where

$$\mathbf{h}_B = \mathbf{H}_B\mathbf{1}_p .$$

Then the conditional likelihood is found to be

$$\begin{aligned} L_1(\mathbf{Y} \mid \omega_B) &= \left(\frac{1}{2\pi}\right)^{\frac{n'}{2}} \frac{2}{\sqrt{2\pi}\theta} \frac{1}{|\Sigma_0|^{1/2}} \\ &\quad \int_0^\infty \exp\left(-\frac{1}{2}(\mathbf{Y} - \delta\mathbf{h}_B)'\Sigma_0^{-1}(\mathbf{Y} - \delta\mathbf{h}_B) - \frac{\delta^2}{2\theta^2}\right) d\delta \\ &= L_0(\mathbf{Y}) 2\sqrt{\frac{1}{\pi(1+\theta^2 d_B)}} \exp\left\{\frac{\theta^2}{2}\frac{(\mathbf{Y}'\Sigma_0^{-1}\mathbf{h}_B)^2}{1+\theta^2 d_B}\right\} \\ &\quad \Phi\left(\frac{\theta\mathbf{Y}'\Sigma_0^{-1}\mathbf{h}_B}{\sqrt{1+\theta^2 d_B}}\right) , \end{aligned}$$

where $\Phi(\ .\ )$ is the cumulative density of a normal distribution and $d_B = \mathbf{h}_B'\Sigma_0^{-1}\mathbf{h}_B$.

For θ small,

$$L_1(\mathbf{Y} \mid \omega_B) = L_0(\mathbf{Y})\{1 + o(\theta)\}\left\{1 + \sqrt{\frac{2}{\pi}}\theta\mathbf{Y}'\Sigma_0^{-1}\mathbf{h}_B + o(\theta)\right\}$$

and if terms of order $o(\theta)$ are neglected, the unconditional likelihood ratio is approximately

$$\frac{L_1(\mathbf{Y})}{L_0(\mathbf{Y})} = 1 + \sqrt{\frac{2}{\pi}}\,\theta\sum_B p(\omega_B)\mathbf{Y}'\Sigma_0^{-1}\mathbf{h}_B\ .$$

Therefore, a statistic to test for one-sided changes at unknown boundary B is seen to be

$$T_n = \sum_B p(\omega_B)\mathbf{Y}'\Sigma_0^{-1}\mathbf{h}_B\ .$$

Woodbury's formula states:

$$(A + CDC')^{-1} = A^{-1} - A^{-1}C(C'A^{-1}C + D^{-1})^{-1}C'A^{-1}\ .$$

Using this result as $\tau \to \infty$, it can be shown that

$$\Sigma_0^{-1} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ .$$

Hence, we have the following theorem:

Theorem 2.2.1 *A Bayes-type statistic for testing the null hypothesis* $H_0 : \delta = 0$ *against the one-sided alternative hypothesis* $H_1 : \delta > 0$ *for some unknown boundary* B *is*

$$T_n = \sum_B p(\omega_B) \left\{ \mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{H}_B \mathbf{1}_p \right\} .$$

The following special cases of Theorem 2.2.1 are considered:

Suppose a set of observations is a sequence, i.e., it is taken from classical regression model

$$Y_j = \sum_{k=0}^{p-1} \beta_k f_k(j/n) + \epsilon_j , \quad j = 1, \ldots, n.$$

Let $p(m)$, $m = 1, 2, \ldots, n-1$, denote the prior probability distribution on the unknown change-point m and $nR_B = \{m+1, \ldots, n\}$. Then T_n can be reduced to

$$T_n = \sum_{m=1}^{n-1} p(m) \left\{ \mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \left(\sum_{k=0}^{p-1} \omega_k X_{mk} \right) \right\} ,$$

where X_{mk} is the k^{th} column vector of the design matrix $\mathbf{X}$ with the first m rows replaced by zeros, which is the statistic derived by Jandhyala and MacNeill (1991).

If an array of observations is taken from a normal distribution with mean β_0 and known variance σ^2, then the design matrix becomes $\mathbf{X} = (1, 1, \ldots, 1)'$ and $\omega_R = \omega_0(B)$. Hence the test statistic for a one-sided change is given by

$$T_n^* = \sum_B p(\omega_B)\omega_B \left\{ \sum_{n^{-1}j \in R_B} (Y_j - \bar{Y}) \right\} .$$

Also, it is consistent with change-point statistic derived by Chernoff and Zacks (1964) when the boundaries reduce to a sequence of change-point at m and $nR_B = \{m+1, \ldots, n\}$.

A more interesting problem is that of testing for two-sided changes in parameters. We have the following theorem:

Theorem 2.2.2 *A Bayes-type statistic for testing the null hypothesis*

$$H_0 : \ \delta_0 = \delta_1 = \cdots = \delta_{p-1} = 0$$

against a two-sided alternative

$$H_1 : \ \delta_k \neq 0$$

for at least some k and for unknown boundary B is given by

$$U_n = \sum_B p(\omega_B)\mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{H}_B\mathbf{H}_B'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \ ,$$

where $\mathbf{H}_B$ is the matrix $\mathbf{H} = (\omega_0(B)f_0(n^{-1}\mathbf{j}), \ldots, \omega_{p-1}(B)f_{p-1}(n^{-1}\mathbf{j}) : \mathbf{j} \leq n\mathbf{1})$ with components replaced by 0 when $n^{-1}\mathbf{j}$ is not in R_B.

Proof. The model under H_0 is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

and under the alternative hypothesis H_1 it is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{H}_B\boldsymbol{\delta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\delta} = (\delta_0, \ldots, \delta_{p-1})'$ and δ_k is the amount of change in the parameter β_k ($k = 0, 1, \ldots, p-1$) at the unknown boundary B. Let the prior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ be as follows:

$$\boldsymbol{\beta} \sim N_p(0, \tau^2\mathbf{I}_p) \text{ and } \boldsymbol{\delta} \sim N_p(0, \theta^2\mathbf{I}_p)$$

with $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ all distributed independently.

Hence

$$(\mathbf{Y} \mid \boldsymbol{\beta}, \omega_B) \sim N_{n^d}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I} + \theta^2 \mathbf{H}_B \mathbf{H}_B') \ ,$$

where $\mathbf{H}_B$ is the matrix $\mathbf{H} = (\omega_0(B) f_0(n^{-1}\mathbf{j}), \ldots, \omega_{p-1}(B) f_{p-1}(n^{-1}\mathbf{j}) : \ \mathbf{j} \leq n\mathbf{1})$ with components replaced by 0 when $n^{-1}\mathbf{j}$ is not in R_B.

We have

$$L_1(\mathbf{Y} \mid \omega_B) = \left(\frac{1}{2\pi}\right)^{\frac{n^d}{2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}\mathbf{Y}'\Sigma^{-1}\mathbf{Y}\right\} \ ,$$

where $\Sigma = \Sigma_0 + \theta^2 D_B$ and $D_B = \mathbf{H}_B \mathbf{H}_B'$.

Therefore

$$\frac{L_1(\mathbf{Y} \mid \omega_B)}{L_0(\mathbf{Y})} \propto \exp\left\{\frac{1}{2}\mathbf{Y}'(\Sigma_0^{-1} - \Sigma^{-1})\mathbf{Y}\right\} \ .$$

As in Jandhyala and MacNeill (1991), to compute Σ^{-1}, we consider

$$\Sigma\Sigma_0^{-1} = \mathbf{I} + \theta^2 D_B \Sigma_0^{-1} \ .$$

Therefore

$$(\Sigma\Sigma_0^{-1})^{-1} = (\mathbf{I} + \theta^2 D_B \Sigma_0^{-1})^{-1} = \mathbf{I} - \theta^2 D_B \Sigma_0^{-1} + o(\theta^2) \ .$$

If one neglects terms of order $o(\theta^2)$ for small θ, one obtains

$$\Sigma_0\Sigma^{-1} = \mathbf{I} - \theta^2 D_B \Sigma_0^{-1} \ .$$

Hence

$$\Sigma_0^{-1} - \Sigma^{-1} \simeq \theta^2 \Sigma_0^{-1} D_B \Sigma_0^{-1} \ .$$

The proof is completed by noting that Woodbury's formula with $\tau \to \infty$ implies that $\Sigma_0^{-1} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

We now consider the special case in which an array of observations is taken from a normal distribution with mean β_0 and known variance σ^2. Then the design matrix becomes $\mathbf{X} = (1, 1, \ldots, 1)$ and $\omega_B = \omega_0(B)$. Hence the two-sided test statistic is given by

$$U_n^* = \sum_B p(\omega_B)\omega_B \left\{ \sum_{n^{-1}j \in R_B} (Y_j - \bar{Y}) \right\}^2 .$$

This is consistent with the change-point test statistic derived by Gardner (1969) when the mean of a sequence of observations changes at m and $nR_B = \{m+1, \ldots, n\}$.

Similar to Theorem 2.2.2, we can derive the Bayes-type statistic for testing the null hypothesis

$$H_0 : \delta = 0$$

against the alternative hypothesis

$$H_1 : \delta \neq 0$$

for some unknown boundary B.

Theorem 2.2.3 *A Bayes-type statistic for testing the null hypothesis* $H_0 : \delta = 0$ *against the two-sided alternative hypothesis* $H_1 : \delta \neq 0$ *for some unknown boundary is*

$$U_n = \sum_B p(\omega_B) \left\{ \mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{H}_B \mathbf{1}_p \mathbf{1}_p' \mathbf{H}_B' (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \right\} .$$

2.3 Test Statistics for Change of Parameters at Some Simple Unknown Boundaries

In this section we illustrate the results of the previous section by considering several simple cases. In particular, we consider a two-dimensional array of n^2 normally distributed observations Y_{ij} $(i,j = 1,\ldots,n)$ and test for the presence of a change in mean at an unknown boundary. This is a special case of the regression model and the unknown lattice boundaries considered in Section 2.2. This is a spatial analogue of the change-point problems discussed by Chernoff and Zacks (1964) and Gardner (1969).

First we consider the collection of boundaries consisting of the rectangles whose left lower corner is $\mathbf{0} = (0,0)$. Such a rectangular boundary requires only another right upper corner $n^{-1}\mathbf{k} = (l/n, k/n)$ to fix the entire boundary. The boundary is denoted by $B_{\mathbf{k}}$, and the lattice points contained by the rectangle, denoted by $R_{B_{\mathbf{k}}}$, are the points $n^{-1}\mathbf{1} \leq n^{-1}\mathbf{j} \leq n^{-1}\mathbf{k}$, where $\mathbf{j} = (j_1, j_2)$, which is illustrated in Figure 2.1.

Using the results of Section 2.2, we can easily obtain:

For the one-sided case,

$$T_{1n} = \sum_{l=1}^{n}\sum_{k=1}^{n} p(\omega_{lk})\omega_{lk}\left\{\sum_{j_1=1}^{l}\sum_{j_2=1}^{k}(Y_{j_1 j_2} - \beta_0)\right\}$$

if β_0 is known, and

$$T_{1n}^{*} = \sum_{l=1}^{n}\sum_{k=1}^{n} p(\omega_{lk})\omega_{lk}\left\{\sum_{j_1=1}^{l}\sum_{j_2=1}^{k}(Y_{j_1 j_2} - \hat{\beta}_0)\right\}$$

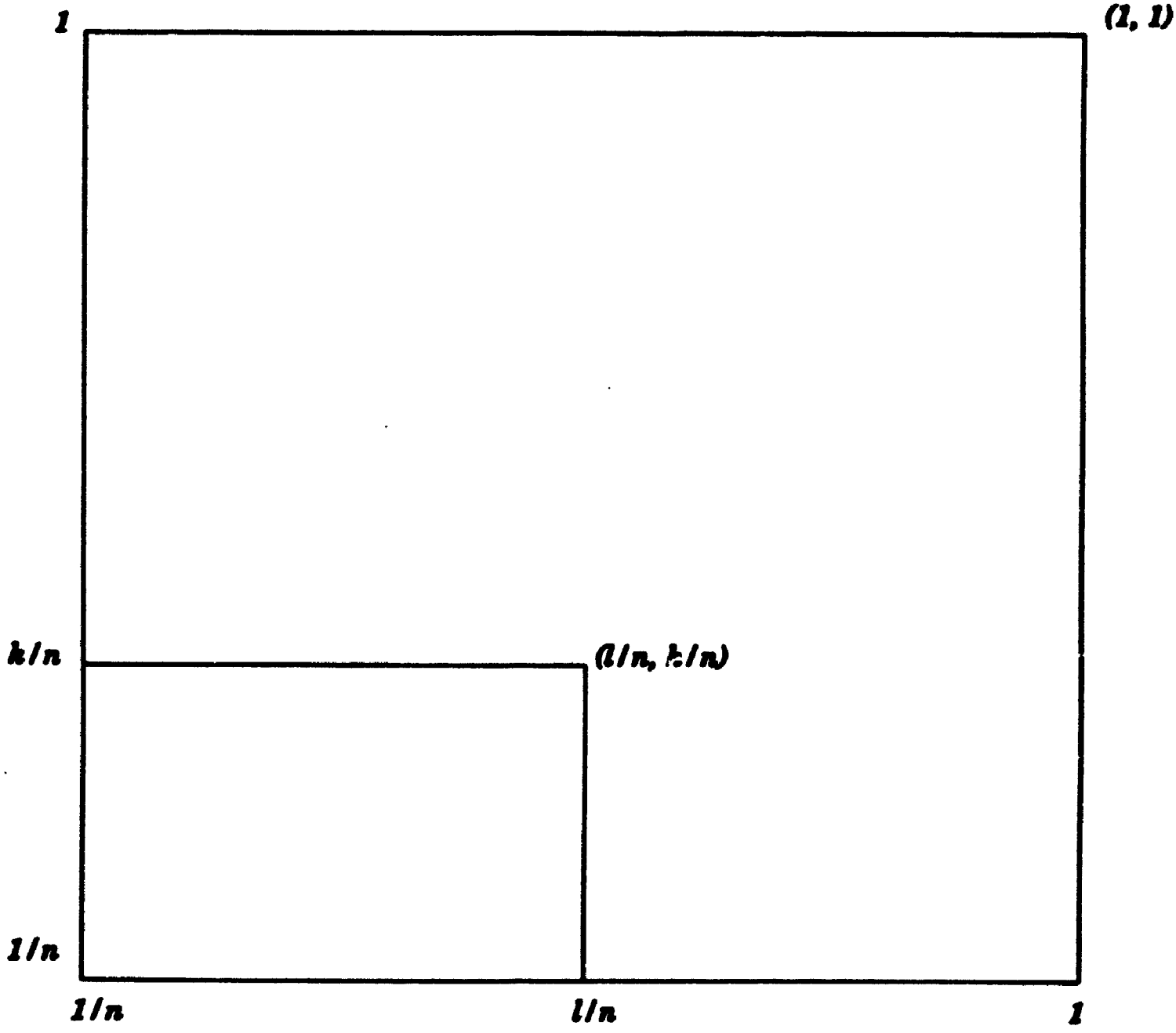

Figure 2.1: Rectangular boundary $\mathcal{B}_k$ with one vertex fixed at 0.

if β_0 is unknown.

For the two-sided case,

$$U_{1n} = \sum_{l=1}^{n}\sum_{k=1}^{n} p(\omega_{lk})\omega_{lk}\left\{\sum_{j_1=1}^{l}\sum_{j_2=1}^{k}(Y_{j_1 j_2} - \beta_0)\right\}^2$$

if β_0 is known, and

$$U_{1n}^{*} = \sum_{l=1}^{n}\sum_{k=1}^{n} p(\omega_{lk})\omega_{lk}\left\{\sum_{j_1=1}^{l}\sum_{j_2=1}^{k}(Y_{j_1 j_2} - \hat{\beta}_0)\right\}^2$$

if β_0 is unknown.

A rectangular boundary with sides parallel to those of the entire space but located at any position within the unit square, requires two points $n^{-1}\mathbf{k} = (l/n, k/n)$ and $n^{-1}\mathbf{r} = (i/n, j/n)$ $(\mathbf{k} \leq \mathbf{r})$ to entirely define the boundary. Such a boundary denoted by $B_{\mathbf{rk}}$ is illustrated in Figure 2.2.

Again, using the results of the previous section, we obtain:

For the one-sided case,

$$T_{2n} = \sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{i=l}^{n}\sum_{j=k}^{n} p(\omega_{lk}^{ij})\omega_{lk}^{ij}\left\{\sum_{j_1=i}^{l}\sum_{j_2=j}^{k}(Y_{j_1 j_2} - \beta_0)\right\}$$

if β_0 is known, and

$$T_{2n}^{*} = \sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{i=l}^{n}\sum_{j=k}^{n} p(\omega_{lk}^{ij})\omega_{lk}^{ij}\left\{\sum_{j_1=i}^{l}\sum_{j_2=j}^{k}(Y_{j_1 j_2} - \hat{\beta}_0)\right\}$$

if β_0 is unknown.

For the two-sided case,

$$U_{2n} = \sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{i=l}^{n}\sum_{j=k}^{n} p(\omega_{lk}^{ij})\omega_{lk}^{ij}\left\{\sum_{j_1=i}^{l}\sum_{j_2=j}^{k}(Y_{j_1 j_2} - \beta_0)\right\}^2$$

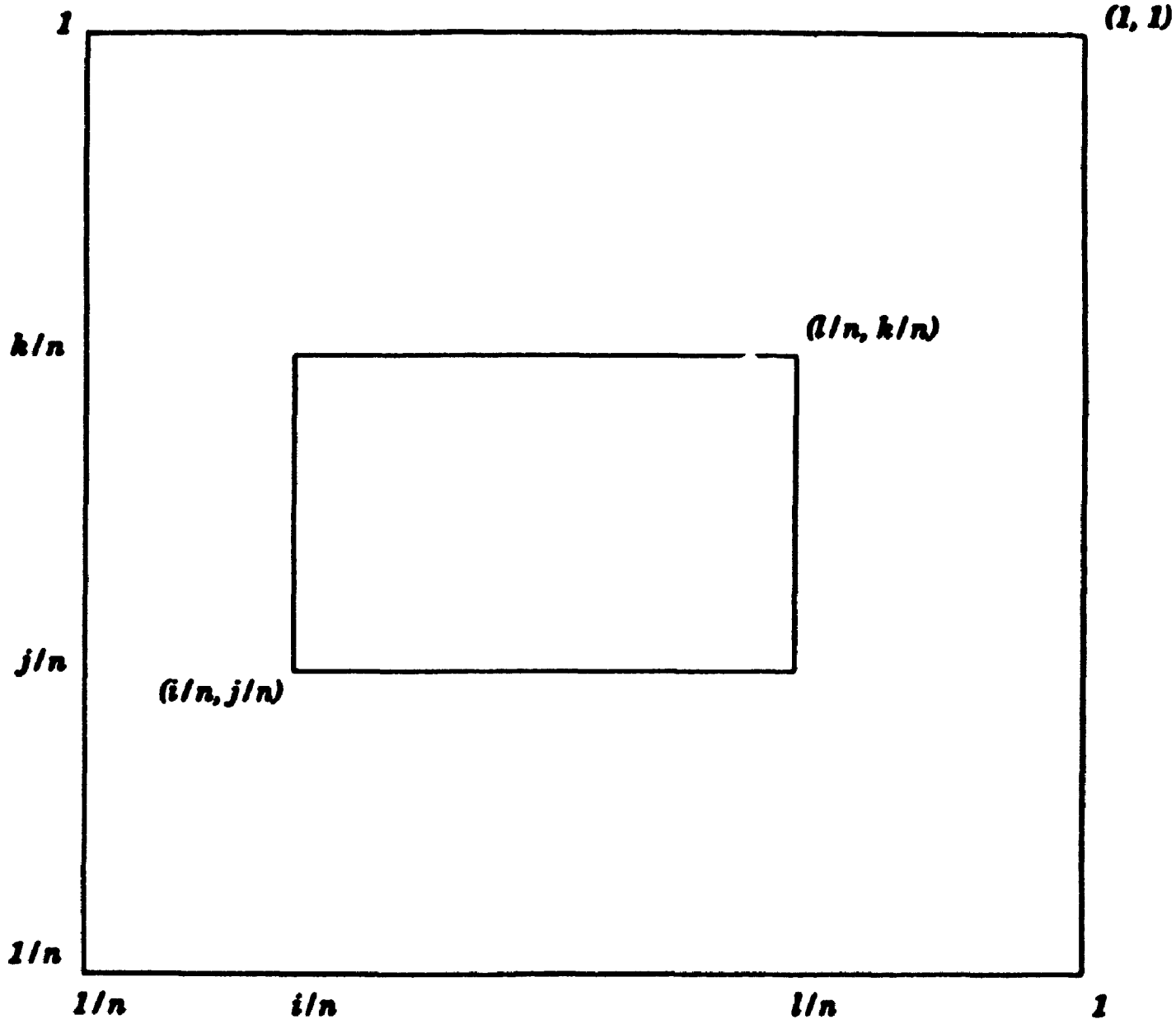

Figure 2.2: Rectangular boundary B_{rk} with sides parallel to the edges of the unit square.

if β_0 is known, and

$$U_{2n} = \sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{i=l}^{n}\sum_{j=k}^{n} p(\omega_{lk}^{ij})\omega_{lk}^{ij}\left\{\sum_{j_1=i}^{l}\sum_{j_2=j}^{k}(Y_{j_1 j_2} - \hat{\beta}_0)\right\}^2$$

if β_0 is unknown.

Similar to the rectangular boundary case, we consider the collection of boundaries consisting of the circles or segments of circles whose centre point $\mathbf{c} = (i/n, j/n)$ is located at any position within the unit square. Such a circular boundary requires a radius r/n $(r \leq n)$ to entirely define the boundary. We consider the collection $\mathbf{S_3}$ of boundaries $B(\mathbf{c}, r)$ such that $(B(\mathbf{c}, r) \cap$ unit square) is not empty. The lattice points contained by the circle and unit square is denoted by the $R_{B(\mathbf{c},r)}$ (see Figure 2.3).

Using the results of Section 2.2, we can easily obtain:

For the one-sided case,

$$T_{3n} = \sum_{\mathbf{S_3}} p(\omega_{B(\mathbf{c},r)})\omega_{B(\mathbf{c},r)}\left\{\sum_{(i,j)\in R_{B(\mathbf{c},r)}}(Y_{ij} - \beta_0)\right\}$$

if β_0 is known, and

$$T_{3n}^{*} = \sum_{\mathbf{S_3}} p(\omega_{B(\mathbf{c},r)})\omega_{B(\mathbf{c},r)}\left\{\sum_{(i,j)\in R_{B(\mathbf{c},r)}}(Y_{ij} - \hat{\beta}_0)\right\}$$

if β_0 is unknown.

For the two-sided case,

$$U_{3n} = \sum_{\mathbf{S_3}} p(\omega_{B(\mathbf{c},r)})\omega_{B(\mathbf{c},r)}\left\{\sum_{(i,j)\in R_{B(\mathbf{c},r)}}(Y_{ij} - \beta_0)\right\}^2$$

if β_0 is known, and

$$U_{3n}^{*} = \sum_{\mathbf{S_3}} p(\omega_{B(\mathbf{c},r)})\omega_{B(\mathbf{c},r)}\left\{\sum_{(i,j)\in R_{B(\mathbf{c},r)}}(Y_{ij} - \hat{\beta}_0)\right\}^2$$

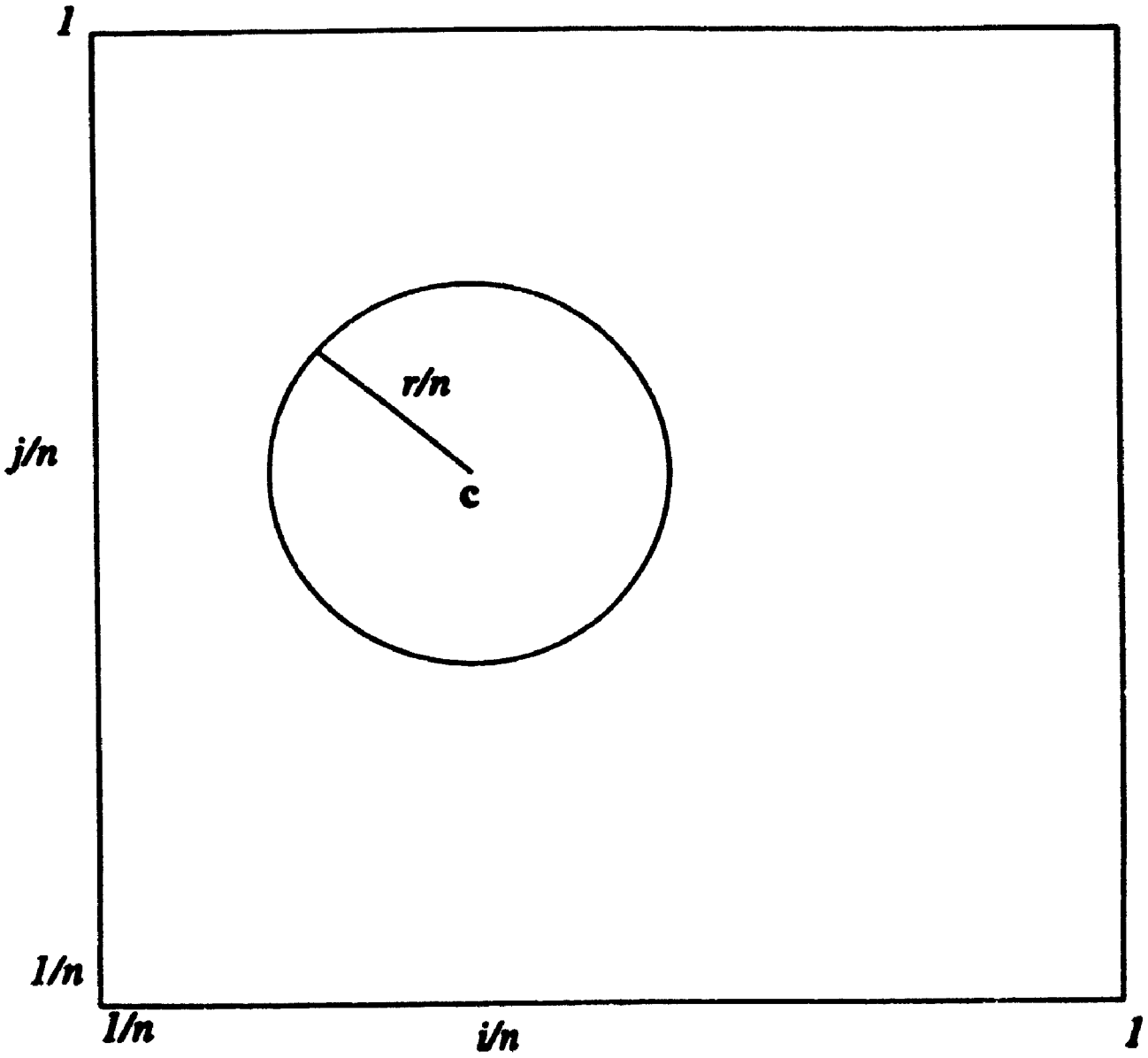

Figure 2.3: Circular boundary $B(\mathbf{c}, r)$.

if β_0 is unknown.

We next consider the case where a boundary is the hypotenuse of an isosceles right triangle whose vertex is $\mathbf{0} = (0,0)$. Such a boundary requires only one other point on a side of the space to be entirely defined. The boundary is denoted by $B^{(l)}$ (see Figure 2.4). Such a boundary divides the space into an isosceles right triangle and a quadrilateral.

Using the result of Section 2.2, we can obtain:

$$U_{4n}^{\bullet} = n\sum_{l=1}^{n} p(\omega_{B^{(l)}})\omega_{B^{(l)}} \left\{ \sum_{j_1=1}^{l} \sum_{j_2=1}^{l-j_1+1} (Y_{j_1 j_2} - \hat{\beta}_0) \right\}^2 + n\sum_{k=1}^{n} p(\omega_{B^{(k)}})\omega_{B^{(k)}} \left\{ \left(\sum_{j_1=1}^{n}\sum_{j_2=1}^{n} - \sum_{j_1=k}^{n}\sum_{j_2=n+k-j_1}^{n} \right) (Y_{j_1 j_2} - \hat{\beta}_0) \right\}^2$$

if β_0 is unknown.

For the general regression model

$$Y_{ij} = \sum_{k=0}^{p-1} \beta_k f_k(i/n, j/n) + \varepsilon_{ij} \ ,$$

using the results of Section 2.2, the test statistics for detecting changes in parameters at unknown boundary as in Figure 2.1 are:

For the one-sided case,

$$T_{1n}^{\bullet} = \sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{i=0}^{p-1} p(\omega_{lk})\omega_{lk} \sum_{j_1=1}^{l}\sum_{j_2=1}^{k} f_i(j_1/n, j_2/n) r_{j_1 j_2} \ ,$$

where $r_{j_1 j_2}$ is regression residual.

For the two-sided case,

$$U_{1n}^{\bullet} = \sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{i=0}^{p-1} p(\omega_{lk})\omega_{lk} \left(\sum_{j_1=1}^{l}\sum_{j_2=1}^{k} f_i(j_1/n, j_2/n) r_{j_1 j_2} \right)^2 \ .$$

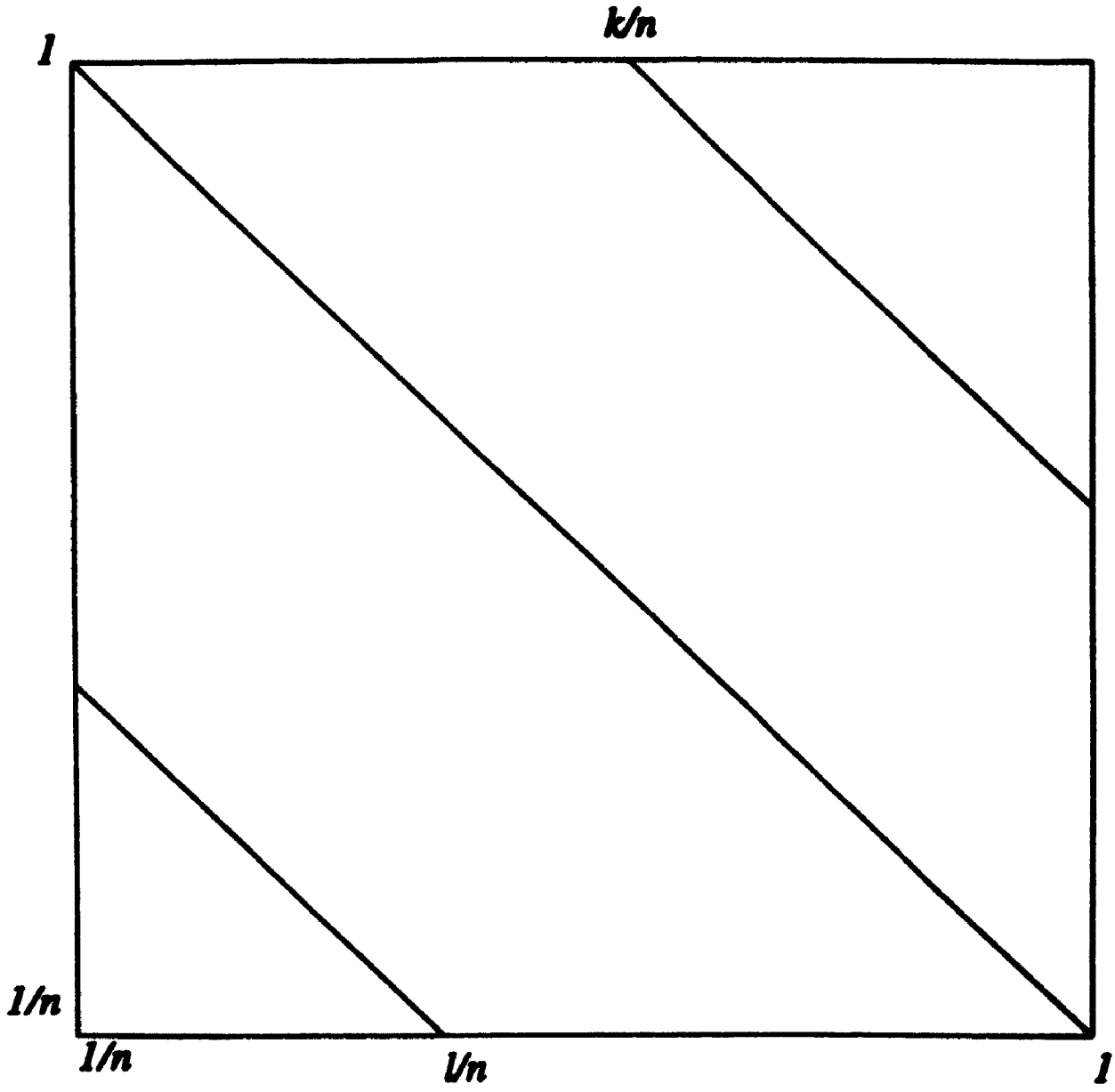

Figure 2.4: Isosceles right triangle or polygon boundary $B^{(l)}$.

This is a spatial analogue of the change-point problems for regression models discussed by Jandhyala and MacNeill (1991).

It is straightforward to extend these results to the d-dimensional solid boundary case. Other collections of boundaries may be considered; these result in test statistics consisting of other functionals of the partial sums of residuals.

2.4 Conclusion

We have derived Bayes-type statistics for testing for the presence of boundaries in spatial data. First, we assume appropriate prior distributions on nuisance parameters associated with both the null and alternative hypotheses. Next, the respective unconditional likelihood functions are obtained through elimination of the nuisance parameters. Finally, a statistic for testing for changes in parameters is derived by the likelihood ratio method. This approach was introduced by Chernoff and Zacks (1964) and developed for regression problems by Jandhyala and MacNeill (1991). We have extended this method to the case of spatial data.

Chapter 3

LOCALLY BEST STATISTICS FOR TESTING FOR THE PRESENCE OF A CHANGE BOUNDARY

3.1 Introduction

Nabeya and Tanaka (1988) discussed the problem of testing for the constancy of regression coefficients under a random walk alternative and obtained a locally best (LB) invariant statistic. Jandhyala and MacNeill (1992) showed that the locally best statistic to test for the constancy of regression coefficients under a random walk and a Bayes-type statistic derived under a change-point alternative are identical. Nyblom and Mäkeläinen (1983) and Nyblom (1989) proposed the LB test for detecting parameter changes under the change-point as well as the random-walk alternative.

In this chapter we discuss a locally optimal level α test for the change-boundary problem. The statistics for testing parameter changes at unknown boundaries are derived by first eliminating the nuisance parameters by the Bayesian method and then dealing with the change parameters through a locally optimal approach.

We recall the generalized Neyman-Pearson lemma, which is the fundamental theorem of statistical inference for testing hypotheses, and review the locally best (LB)

test for the single parameter (available in standard advanced inference books) and the multiparameter case (SenGupta and Vermeire (1986)) in Section 3.2. In Section 3.3 we derive statistics for testing an array of normal observations for no change in mean against alternatives involving changes at unknown boundary. In section 3.4 we discuss the case where the observations are taken from a regression model.

3.2 Locally Best Test

To begin we recall the generalized Neyman-Pearson fundamental lemma:

Lemma 3.1 *Let $f_0, f_1, \ldots, f_n$ be integrable functions, $0 \leq \phi \leq 1$, and*

$$\int \phi(x) f_i(x) dx = c_i \text{ (given)}, \ i = 1, 2, \ldots, n.$$

If there exists a function $\phi_0(.)$ such that

$$(a) \quad \int \phi_0(x) f_i(x) dx = c_i, \ i = 1, 2, \ldots, n.$$

$$(b) \quad \phi_0(x) = \begin{cases} 1 & \text{if } f_0(x) > \sum_{i=1}^n k_i f_i(x) \\ \gamma(x) & \text{if } f_0(x) = \sum_{i=1}^n k_i f_i(x) \\ 0 & \text{if } f_0(x) < \sum_{i=1}^n k_i f_i(x) \end{cases}$$

where $0 \leq \gamma(x) \leq 1$, and $k_i \geq 0$, $i = 1, 2, \ldots, n$. Then

$$\int \phi_0(x) f_0(x) dx \geq \int \phi(x) f_0(x) dx \ .$$

We review the locally best test (Rao (1973)) as follows. We let $f(x, \theta)$ be a one-parameter density function for the random variable X and consider the null hypothesis $H_0 : \ \theta = \theta_0$. We wish to obtain critical regions which are best for alternatives close to the null hypothesis and hope that such regions will also do well for distant

alternatives. Such tests are said to be locally best (LB). Assume that the distribution of the random observation X is such that the power function

$$\beta_\phi(\theta) = E[\phi(X)] = \int \phi(x) f(x|\theta) dx$$

of any test ϕ admits a Taylor expansion

$$\beta_\phi(\theta) = \alpha + (\theta - \theta_0)\beta'_\phi(\theta_0) + \frac{(\theta - \theta_0)^2}{2!}\beta''_\phi(\theta_0) + \cdots ,$$

where $\beta_\phi(\theta_0) = \alpha$. If we assume that integration and differentiation can be interchanged, then we can obtain a LB one-sided test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ by maximizing $\beta'_\phi(\theta)$ at θ_0. To apply the Neyman-Pearson Lemma we have the following theorem for a LB one-sided test:

Theorem 3.1 *A critical region of the LB test for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ is formed by the subset of the sample space such that*

$$\frac{\partial}{\partial\theta} f(x \mid \theta)\Big|_{\theta=\theta_0} > k\, f(x \mid \theta)|_{\theta=\theta_0} ,$$

where k is chosen to satisfy $\beta_\phi(\cdot) = \alpha$.

If the alternative H_1 is two-sided, i.e. $H_1 : \theta \neq \theta_0$, then we maximize $\beta''_\phi(\theta)$ at θ_0 by imposing the local unbiased restriction $\beta'_\phi(\theta_0) = 0$. Such a test is said to be a LB unbiased test. Thus the LB unbiased level α test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is subject to $\beta_\phi(\theta_0) = \alpha$ and $\beta'_\phi(\theta_0) = 0$ and maximizes $\beta''_\phi(\theta)$ at θ_0. We can directly use the generalized Neyman-Pearson lemma to obtain the following theorem.

Theorem 3.2 *A critical region of the LB unbiased level α test $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$ is the subset of the sample space such that*

$$\frac{\partial^2}{\partial^2\theta} f(x \mid \theta)\Big|_{\theta=\theta_0} > k_1 f(x \mid \theta)|_{\theta=\theta_0} + k_2 \frac{\partial}{\partial\theta} f(x \mid \theta)\Big|_{\theta=\theta_0},$$

where k_1 and k_2 are chosen to satisfy $\beta_\phi(\theta_0) = \alpha$ and $\beta'_\phi(\theta_0) = 0$.

In the p-parameter case, SenGupta and Vermeire (1986) proposed a locally most mean powerful unbiased (LMMPU) test for multiparameter hypotheses. This optimality property is a generalization to the multiparameter case of LB unbiased tests.

Let a p-parameter density function be $f(x, \theta)$, where $\theta = (\theta_1, \ldots, \theta_p)'$. Consider the null hypothesis $H_0: \theta = \theta_0$, where $\theta_0 = (\theta_1^0, \ldots, \theta_p^0)'$. The local behavior of the power $\beta(\theta)$ at θ_0 can be studied by its Taylor expansion at θ_0. That is, we have

$$\beta(\theta) = \beta(\theta_0) + (\theta - \theta_0)'\dot{\beta}(\theta_0) + \frac{1}{2!}(\theta - \theta_0)'\ddot{\beta}(\theta_0)(\theta - \theta_0) + \cdots,$$

where $\dot{\beta}(\theta_0)$ and $\ddot{\beta}(\theta_0)$ are the gradient vector and Hessian matrix of $\beta(\theta)$ at θ_0. That is,

$$\dot{\beta}(\theta_0) = \left(\frac{\partial}{\partial\theta_1}\beta(\theta), \ldots, \frac{\partial}{\partial\theta_p}\beta(\theta)\right)'\Big|_{\theta=\theta_0}$$

and

$$\ddot{\beta}(\theta_0) = \left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\beta(\theta)\right)\Big|_{\theta=\theta_0}.$$

This kind of test maximizes the average power on a p-dimensional spherical neighborhood of the null hypothesis among the locally unbiased level α tests and its critical regions are easily constructed.

Applying the generalized Neyman-Pearson lemma, we have the following result for the multiparameter case due to SenGupta and Vermeire (1986).

Theorem 3.3 *A critical region of a LMMPU test for* $H_0: \theta = \theta_0$ *against* $H_0: \theta \neq \theta_0$ *can be formed by*

$$\sum_{i=1}^{p} \frac{\partial^2}{\partial^2 \theta_i} f(x \mid \theta)\Big|_{\theta=\theta_0} > k\, f(x \mid \theta)|_{\theta=\theta_0} + \sum_{i=1}^{p} k_i \frac{\partial}{\partial \theta_i} f(x \mid \theta)\Big|_{\theta=\theta_0}$$

where the constants $k, k_1, \ldots, k_p$ *satisfy the conditions (1)* $\beta(\theta_0) = \alpha$, *(2)* $\dot{\beta}(\theta_0) = 0$ *and (3)* $\ddot{\beta}(\theta_0)$ *is nonnegative definite.*

3.3 Detection of Mean Change at Unknown Boundaries from Normal Data

As indicated above, Chernoff and Zacks (1964) proposed a one-sided Bayes-type test statistic for change of parameter at unknown times in the mean of a sequence of independent normal random variables and this approach was adapted by Gardner (1969) for two-sided tests of parameter changes in a sequence of normal random variables. In this section, we derive statistics for testing for the presence of boundaries in spatial data. First, we apply a Bayesian procedure introduced to the change-point problem by Chernoff and Zacks (1964). This procedure places convenient prior distributions on the nuisance parameters (if they exist) and then eliminates them through integration. Next, a prior distribution on the set of possible change-boundaries is used to obtain the distribution of the data conditional only upon the possible changes in mean. Finally, a statistic for testing for changes in mean is obtained in the LB framework.

Let $\{Y_j : j \in Z_+^d\}$ be an array of n^d observations taken from a normal distribution

$N(\mu_j, \sigma^2)$. We wish to test the hypothesis of no change in mean, i.e, $H_0: \mu_j = \mu$ for all $j \leq n1$, against certain alternative changes in mean at some unknown boundary B. To specify alternatives we let δ be the amount of change and ω_B be 1 or 0 according to whether there is or is not change at unknown boundary B.

Hence the null and alternative hypotheses may be written as follows:

$$H_0: \delta = 0$$

against

$$H_1: \delta > 0$$

for some unknown boundary B.

We first consider the case where μ is known and later discuss the case where μ is unknown.

If μ is known (set $\mu = 0$), then under the alternative hypothesis H_1,

$$(\mathbf{Y} \mid \delta, \omega_B) \sim N_{n^d}(\delta\omega_B \mathbf{1}_B, \mathbf{I}),$$

where $\mathbf{Y} = \{Y_j: 1 \leq j \leq n1\}$ and $\mathbf{1}_B$ is an $n^d \times 1$ vector of ones $1'$ with components replaced by 0 when $n^{-1}j$ is not in the region enclosed by B.

Hence the p.d.f. of $(\mathbf{Y} \mid \delta)$ is

$$f(\mathbf{Y} \mid \delta) = \sum_B p(\omega_B) f(\mathbf{Y} \mid \delta, \omega_B),$$

where $p(\omega_B)$ is a prior probability for the unknown boundary B and

$$f(y \mid \delta, \omega_B) = \left(\frac{1}{2\pi}\right)^{\frac{n^d}{2}} \exp\left\{-\frac{1}{2}\sum_{n^{-1}j \in R_B}(Y_j - \delta\omega_B)^2 - \sum_{n^{-1}j \in R_B} Y_j^2\right\},$$

where R_B is a region enclosed by B and $R_{\mathcal{L}}$ is the complementary set of R_B.

The LB test statistic according to Theorem 3.1 is now obtained as

$$\frac{\partial}{\partial\delta} f(\mathbf{Y} \mid \delta)\Big|_{\delta=0} > k\, f(\mathbf{Y} \mid \delta)|_{\delta=0} ,$$

where k is a chosen constant, which yields

$$S = \sum_B p(\omega_B)\omega_B \sum_{n^{-1}\mathbf{j}\in R_B} Y_{\mathbf{j}} ,$$

as the test statistic.

When μ is unknown, assume $\mu \sim N(0, \tau^2)$. Then under the alternative hypothesis H_1,

$$(\mathbf{Y} \mid \delta, \omega_B) \sim N_{n^d}(\delta\omega_B \mathbf{1}_B, \Sigma_0) ,$$

where $\Sigma_0 = \mathbf{I} + \tau^2 \mathbf{1}\mathbf{1}'$.

Since Σ_0 is a positive definite matrix, there exists a full rank matrix C such that $C\Sigma_0 C' = \mathbf{I}$, i.e., $\Sigma_0^{-1} = C'C$.

Let $\mathbf{Z} = C\mathbf{Y}$, then

$$(\mathbf{Z} \mid \delta, \omega_B) \sim N_{n^d}(\delta\omega_B C\mathbf{1}_B, \mathbf{I}) .$$

If we use the same argument as above, we obtain the test statistic

$$S = \sum_B p(\omega_B)\omega_B \mathbf{Y}'\Sigma_0^{-1}\mathbf{1}_B .$$

By Woodbury's formula with $\tau \to \infty$ it can be shown that $\Sigma_0^{-1} = \mathbf{I} - \frac{1}{n^d}\mathbf{1}\mathbf{1}'$.

Hence

$$S = \sum_B p(\omega_B)\omega_B \sum_{n^{-1}\mathbf{j}\in R_B} (Y_{\mathbf{j}} - \bar{Y}) ,$$

where $\bar{Y} = \frac{1}{n^2} \sum_{j \le n1} Y_j$.

For the two-sided case, the null and alternative hypotheses are

$$H_0 : \delta = 0$$

against

$$H_1 : \delta \neq 0$$

for some unknown boundary B.

The LB unbiased test statistics can be obtained as follows:

$$\frac{\partial^2}{\partial^2 \delta} f(\mathbf{Z} \mid \delta)\Big|_{\delta=0} > k_1 f(\mathbf{Z} \mid \delta)|_{\delta=0} + k_2 \frac{\partial}{\partial \delta} f(\mathbf{Z} \mid \delta)\Big|_{\delta=0},$$

where k_1 and k_2 are chosen constants to satisfy the LB unbiased criteria (see Theorem 3.2).

Since the distribution is symmetric, then $k_2 = 0$. Hence, we have:

When μ (assume $\mu = 0$) is known, the test statistic is

$$S = \sum_B p(\omega_B)\omega_B \sum_{n^{-1}j \in R_B} Y_j^2 .$$

When μ is unknown, the test statistic is

$$S^* = \sum_B p(\omega_B)\omega_B \sum_{n^{-1}j \in R_B} (Y_j - \bar{Y})^2 .$$

3.4 Derivation of the Statistics for Detecting Changes in Regression Parameters at Unknown Boundaries

In Chapter 2, the Bayes-type statistics for detecting changes in regression parameters at unknown boundaries are derived. In this section, we derive statistics for testing

for the presence of boundaries in spatial data in the LB (or LMMPU) framework by first eliminating the regression coefficient nuisance parameters by the Bayesian method and then dealing with the change parameters through this LB (or LMMPU) approach.

With the notation of Chapter 2, the model under the alternative hypothesis is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{H}_B \mathbf{1}_p \delta + \boldsymbol{\epsilon} \ ,$$

where δ denotes the magnitude of the one-sided change common to all the regression coefficient paremeters.

Hence the null and alternative hypotheses are

$$H_0 : \ \delta = 0$$

against

$$H_1 : \ \delta > 0$$

for some unknown boundary B.

If $\boldsymbol{\beta} \sim N(\mathbf{0}, \tau^2 \mathbf{I}_p)$, then under the alternative hypothesis H_1,

$$(\mathbf{Y} \mid \delta, \omega_B) \ \sim \ N_{n^d}(\mathbf{H}_B \mathbf{1}_p \delta, \Sigma_0) \ ,$$

where $\Sigma_0 = \mathbf{I} + \tau^2 \mathbf{X}\mathbf{X}'$.

Since Σ_0 is a positive definite matrix, there exists a full rank matrix C such that $C\Sigma_0 C' = \mathbf{I}$, i.e. $\Sigma_0^{-1} = C'C$.

Let $\mathbf{Z} = C\mathbf{Y}$, then

$$(\mathbf{Z} \mid \delta, \omega_B) \sim N_{n^d}(C\mathbf{H}_B \mathbf{1}_p \delta, \mathbf{I}) \ .$$

The p.d.f. of $(\mathbf{Z} \mid \delta)$ is

$$f(\mathbf{Z} \mid \delta) = \sum_B p(\omega_B) f(\mathbf{Z} \mid \delta, \omega_B)$$

where

$$f(\mathbf{Z} \mid \delta, \omega_B) = \left(\frac{1}{2\pi}\right)^{\frac{nd}{2}} \exp\left\{-\frac{1}{2}(\mathbf{Z} - C\mathbf{H}_B\mathbf{1}_p\delta)'(\mathbf{Z} - C\mathbf{H}_B\mathbf{1}_p\delta)\right\} .$$

The LB test statistic according to Theorem 3.1 is now obtained as

$$\frac{\partial}{\partial\delta} f(\mathbf{Z} \mid \delta)\Big|_{\delta=0} > k\ f(\mathbf{Z} \mid \delta)|_{\delta=0} ,$$

where k is a chosen constant, which yields

$$S = \sum_B p(\omega_B)\mathbf{Y}'\Sigma_0^{-1}\mathbf{H}_B\mathbf{1}_p .$$

By Woodbury's formula with $\tau \to \infty$ it can be shown that $\Sigma_0^{-1} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The statistic S is identical to T_n in Chapter 2.

For the two-sided case, the null and alternative hypotheses are

$$H_0 : \ \delta = 0$$

against

$$H_1 : \ \delta \neq 0$$

for some unknown boundary B.

The LB unbiased test statistic can be obtained as follows:

$$\frac{\partial^2}{\partial^2\delta} f(\mathbf{Z} \mid \delta)\Big|_{\delta=0} > k_1 f(\mathbf{Z} \mid \delta)|_{\delta=0} + k_2 \frac{\partial}{\partial\delta} f(\mathbf{Z} \mid \delta)\Big|_{\delta=0} ,$$

where k_1 and k_2 are chosen constants to satisfy the LB unbiased criteria (see Theorem 3.2). Since the distribution is symmetric, then $k_2 = 0$.

Hence, we obtain

$$S^* = \sum_B p(\omega_B) \left\{ \mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{H}_B \mathbf{1}_p \mathbf{1}_p' \mathbf{H}_B'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \right\} ,$$

as the test statistic. Again, this is the same as the test statistic in Chapter 2.

We make the following observations regarding the more realistic case when the changes in the regression coefficient parameters are not necessarily equal. That is, the model under the alternative hypothesis is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{H}_B\boldsymbol{\delta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\delta} = (\delta_0, \ldots, \delta_{p-1})'$ and δ_i is the amount of change in the parameter β_i ($i = 0, 1, \ldots, p-1$) at the unknown boundary B, the null and alternative hypotheses are

$$H_0 : \ \delta_0 = \cdots = \delta_{p-1} = 0$$

against

$$H_1 : \ \delta_i \neq 0$$

for at least some i and unknown boundary B.

The p.d.f. of $(\mathbf{Z} \mid \boldsymbol{\delta})$ is

$$f(\mathbf{Z} \mid \boldsymbol{\delta}) = \sum_B p(\omega_B) f(\mathbf{Z} \mid \boldsymbol{\delta}, \omega_B)$$

where

$$f(\mathbf{Z} \mid \boldsymbol{\delta}, \omega_B) = \left(\frac{1}{2\pi}\right)^{\frac{n^*}{2}} \exp\left\{-\frac{1}{2}(\mathbf{Z} - C\mathbf{H}_B\boldsymbol{\delta})'(\mathbf{Z} - C\mathbf{H}_B\boldsymbol{\delta})\right\} .$$

Applying the SenGupta and Vermeire (1986) procedure for the multiparameter case, the test statistic can be shown to be defined by

$$\sum_{i=0}^{p-1} \frac{\partial^2}{\partial^2 \delta_i} f(\mathbf{Z} \mid \boldsymbol{\delta})\Big|_{\boldsymbol{\delta}=\mathbf{0}} > k\, f(\mathbf{Z} \mid \boldsymbol{\delta})|_{\boldsymbol{\delta}=\mathbf{0}} + \sum_{i=0}^{p-1} k_i \frac{\partial}{\partial \delta_i} f(\mathbf{Z} \mid \boldsymbol{\delta})\Big|_{\boldsymbol{\delta}=\mathbf{0}}$$

where k and k_i $(i = 0, 1, \ldots, p-1)$ are constants to satisfy some conditions (see Theorem 3.3). Since the distribution is symmetric, then $k_0 = k_1 = \cdots = k_{p-1} = 0$. Hence we obtain

$$Q = \sum_{B} p(\omega_B) \mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{H}_B\mathbf{H}_B'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \ .$$

This also is the same the test statistic in Chapter 2.

3.5 Conclusion

We have derived test statistics for detecting a change-boundary in spatial data by the LB (c.f. LMMPU) procedure. This approach first uses a Bayesian method introduced to the change-point problem by Chernoff and Zacks (1964). It places convenient prior distributions on the nuisance parameters and then eliminates them through integration. Next, a prior distribution on the set of possible change-boundaries is used to obtain the distribution of the data conditional only upon the possible changes in the parameters. Finally, a statistic for testing for changes in parameters is obtained by a locally best (or locally most mean unbiased) procedure. The test statistics derived in this chapter are the same as those of Chapter 2. Hence, the Bayes-type tests possess a local optimality property.

Chapter 4

LIMIT DISTRIBUTION THEORY

4.1 Introduction

Residual processes defined by partial sums of regression residuals were first derived by MacNeill (1978a,b). Jandhyala and MacNeill (1989, 1991) and Jandhyala and Minogue (1993) used properties of residual processes to obtain asymptotic distributions for partial sums of linear functions of regression residuals. Tang and MacNeill (1993) derived residual processes for stationary time series.

In this chapter, we discuss set indexed partial sums of regression residuals for spatial data. We review weak convergence results for set indexed partial sum processes in Section 2. We indicate how weak convergence for set indexed partial sum processes can be applied to obtain large sample theory for set indexed partial sums of regression residuals and large sample distributions for selected test statistics.

4.2 Set indexed Partial Sum Processes

Because our change-boundary statistics are defined in terms of set indexed partial sums, we review some aspects of the theory for such sums prior to consideration

of large sample distributional results for our statistics. We consider an array of real valued random variables indexed by the d-dimensional positive integer lattice, $\{Y_{\mathbf{j}} : \mathbf{j} = (j_1, j_2, \ldots, j_d)\}$.

Pyke (1973, 1983), Bass and Pyke (1984) and Alexander and Pyke (1986) developed a methodology for the limits of set indexed partial sum processes. In this methodology for any bounded $B \in \mathcal{B}$, the class of all Borel sets in $\mathcal{R}^d$,

$$S(B) = \sum_{\mathbf{j} \in B} Y_{\mathbf{j}} .$$

Let $\mathcal{A}$ be a family of Borel subsets of the unit cube $I^d := [0,1]^d$ and define the normalized partial sum process by

$$S_n(A) = n^{-d/2} S(nA) ;$$

where $nA = \{nx : x \in A\}$. To avoid difficulties that arise when the lattice points in a set A are not in some sense representative of A, it is necessary to consider an appropriate smoothed version of the partial sum process as follows. For $B \in \mathcal{B}$, define

$$Y(B) = \sum_{\mathbf{j}} |B \cap C_{\mathbf{j}}| Y_{\mathbf{j}} ,$$

where $C_{\mathbf{j}}$ denotes the unit cube $(\ \mathbf{j} - \mathbf{1},\ \mathbf{j}\ ]$, $|\ .\ |$ denotes Lebegue measure and $\mathbf{1} = (1, 1, \ldots, 1) \in \mathcal{Z}_+^d$. The appropriate smoothed partial sum process indexed by $\mathcal{A}$ is then given by

$$Z_n(A) = n^{-d/2} Y(nA) = n^{-d/2} \sum_{\mathbf{j}} b_{n\mathbf{j}}(A) Y_{\mathbf{j}}$$

for $A \in \mathcal{A}$, where

$$b_{n\mathbf{j}}(A) = |(nA) \cap C_{\mathbf{j}}| .$$

The Brownian process Z indexed by $\mathcal{A}$ is defined as follows:

$Z = \{Z(A) : A \in \mathcal{A}\}$ is a Gaussian process with zero mean and

$$cov(Z(A_1), Z(A_2)) = |A_1 \cap A_2| \ ,$$

$A_1, A_2 \in \mathcal{A}$. Then we have the following results due to Alexander and Pyke (1986).

Theorem 4.2.1 *If $\{Y_j : j \in Z_+^d\}$ are iid with mean zero and variance 1, then the finite-dimensional distributions of $\{Z_n(B) : B \in \mathcal{B}^d \cap I^d\}$ converge weakly to those of the Brownian process $\{Z(B) : B \in \mathcal{B}^d \cap I^d\}$, where $\mathcal{B}^d$ is the class of all Borel sets in $\mathcal{R}^d$.*

Furthermore, if $\mathcal{A}$ satisfies certain mild conditions, tightness can be established hence showing that the sequence of stochastic processes Z_n converges weakly to Z (see Pyke (1983), Bass and Pyke (1984) and Alexander and Pyke (1986)).

In application, it is sometimes not convenient to use set indexed partial sums with smoothing operators. However, if the sets in $\mathcal{A}$ have structural restrictions on their boundaries, such as is true for convex sets, then the smoothing operator can be dropped (Alexander and Pyke (1986)). That is, if A in $\mathcal{A}$ is a convex set, then the partial sum becomes $Z_n(A) = n^{-d/2} \sum_{j \in nA} X_j$, and it converges weakly to the Brownian process. We denote the class of convex sets in $\mathcal{A}$ as $\mathcal{A}^*$.

Pyke (1973) discussed the partial sums of matrix arrays, which are special cases for sets in $\mathcal{A}^*$. Let K_d be the set of d-tuples $k = (k_1, k_2, \ldots, k_d)$ with positive integers for coordinates and let $\leq$ denote the coordinate-wise partial ordering on K_d. If

$\{X_j : j \in K_d\}$ is an iid r.v. of matrix arrays with mean zero and variance 1, the partial sum can be defined as

$$S_k = \sum_{j \le k} X_j , \quad k \in K_d .$$

Set $Z_n(t) = n^{-d/2} S_{[nt]}$, where $t \in I^d$ and $[\ .\ ]$ is the 'greatest-integer-contained-in' function defined coordinate wise. It can be shown that Z_n converges weakly to Z (see Pyke (1973)), where Z is a Brownian sheet; that is, a Gaussian process with mean zero and $cov(Z(s), Z(t)) = |s \wedge t|$, $s,t \in I^d$, where $s \wedge t = (s_1 \wedge t_1) \ldots (s_d \wedge t_d)$ and $s_1 \wedge t_1 = min(s_1, t_1)$.

4.3 Distribution of Test Statistics

Application of the statistics derived in the previous sections requires that their distribution theory be available. We note that the one-sided statistics are linear forms in independent random variables and the two-sided statistics are quadratic forms.

The one-sided statistics will be distributed normally for finite samples. This would be true asymptotically when the noise variables are non-normal but are iid with mean zero and variance σ^2. If σ^2 is unknown, as is usually the case, then one can obtain a statistic by replacing σ^2 with a consistent estimate, and the large sample distributions remain unchanged.

Distribution theory for the quadratic forms representing two-sided test statistics is more complicated. We now pursue the problem of detecting the presence of possible boundaries at unknown location without making any a priori assumptions on the

location of the possible boundaries. Our approach will be to develop statistics based on set indexed partial sums of regression residuals.

To begin we consider n^d observations taken on a regular lattice on the unit cube $I^d := [0,1]^d$; see Figure 2.1 for the two-dimensional case. The observation at the point

$$\frac{1}{n}\mathbf{j} = \left(\frac{j_1}{n}, \ldots, \frac{j_d}{n}\right)$$

is denoted by $Y_{\mathbf{j}}$ and $Y_{\mathbf{j}} \sim N(\mu_{\mathbf{j}}, \sigma^2)$. As defined in Section 4.2, partial sums indexed by $\mathcal{A}^*$, the convex subsets of the unit cube I^d, are given by

$$S(A) = \sum_{\mathbf{j} \in nA} (Y_{\mathbf{j}} - \mu_{\mathbf{j}}) \ , \quad A \in \mathcal{A}^* \ ,$$

where $nA = \{nx : x \in A\}$. It can be shown (Alexander and Pyke (1986)) that

$$\frac{1}{n^{d/2}\sigma} S(A) \Rightarrow Z(A) \ , A \in \mathcal{A}^* \ ,$$

where Z is a Brownian process; convergence is in the sense of weak convergence of probability measures.

Furthermore, if

$$\bar{Y} = \frac{1}{n^d} \sum_{\mathbf{j} \le n\mathbf{1}} Y_{\mathbf{j}} \quad \text{and} \quad S^*(A) = \sum_{\mathbf{j} \in nA} (Y_{\mathbf{j}} - \bar{Y}) \ ,$$

then

$$\frac{1}{n^{d/2}\sigma} S^*(A) \Rightarrow Z_0(A) \ , A \in \mathcal{A}^* \ ,$$

where Z_0 is a Brownian bridge process with

$$E[Z_0(A)] = E[Z_0(I^d)] = 0 \ , \quad A \in \mathcal{A}^*$$

and

$$E[Z_0(A)Z_0(B)] = |A \cap B| - |A||B| \ , \quad A, B \in \mathcal{A}^* \ ,$$

where $|\ .\ |$ denotes Lebegue measure. The relationship between the Brownian process and the Brownian bridge process indexed by $\mathcal{A}$, the class of Borel subsets in I^d, is as follows:

$$Z_0(A) = Z(A) - |A|Z(I^d) \ , \quad A \in \mathcal{A} \ .$$

Now consider a general regression model. We let $f_k(.,\ldots,.)$, $(k = 0, 1, \ldots, p-1)$ be a set of d-variate regression functions, and let $\{\epsilon_{\mathbf{j}} : \mathbf{j} \leq n\mathbf{1}\}$ be an array of iid real-valued r.v. from $N(0, \sigma^2)$. Then we define an array of dependent variables $\{Y_{\mathbf{j}} : \mathbf{j} \leq n\mathbf{1}\}$ on a regular lattice on the d-dimensional unit cube I^d as follows:

$$Y_{\mathbf{j}} = \sum_{k=0}^{p-1} \beta_k f_k(n^{-1}\mathbf{j}) + \epsilon_{\mathbf{j}}$$

where $\mathbf{j} = (j_1, \ldots, j_d)$ is the lattice of positive integer.

If we denote the vector of regression coefficients by $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$, the design matrix by $\mathbf{X}$, the vector of observations by $\mathbf{Y}$ and the vector of noise variables by $\boldsymbol{\epsilon}$, then the model may be written in matrix form as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{(stacked)}.$$

The least squares estimator for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}$ where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \ .$$

We establish the following notation: $\mathbf{t}_d \equiv (t_1, \ldots, t_d)$ and $d\mathbf{t}_d \equiv dt_1 \cdots dt_d$. The vector of regressor functions evaluated at $\mathbf{t}_d$ is denoted by $\mathbf{f}(\mathbf{t}_d)$.

As $n \to \infty$, $\frac{1}{n^2}(\mathbf{X}'\mathbf{X})$ has as its $(r_1, r_2)^{th}$ component

$$\int_0^1 \cdots \int_0^1 f_{r_1}(t_d) f_{r_2}(t_d) dt_d \ .$$

We define the matrix G as follows:

$$\lim_{n\to\infty} n^d (\mathbf{X}'\mathbf{X})^{-1} = \mathbf{G} \ .$$

The inverse exists provided the regressor functions are linearly independent.

For any $A \in \mathcal{A}^*$, the convex subsets of the unit cube I^d, let e_{nA} denote the n^d-dimensional stacked vector that has components equal to 1 if $n^{-1}\mathbf{j} \in A$ and equal to 0 otherwise. The partial sums of residuals can then be denoted by

$$\mathbf{e}'_{nA}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\epsilon \ .$$

Then

$$\mathbf{e}'_{nA}\mathbf{I}\epsilon = \sum_{\mathbf{j}\in nA} \epsilon_{\mathbf{j}} \ .$$

Also, it may be shown that:

$$\frac{1}{n^{d/2}\sigma}\mathbf{e}'_{nA}\mathbf{I}\epsilon \Rightarrow Z(A) \ ,$$

$$\frac{1}{n^{d/2}}\mathbf{e}'_{nA}\mathbf{X} \to \underbrace{\int \cdots \int}_{A} \mathbf{f}(\mathbf{s}_d) d\mathbf{s}_d \ ,$$

where $\mathbf{f}(\mathbf{s}_d)$ is the vector of regressor functions, and

$$\frac{1}{n^{d/2}\sigma}\mathbf{X}'\epsilon \Rightarrow \int_0^1 \cdots \int_0^1 \mathbf{f}(\mathbf{s}_d) dZ(\mathbf{s}_d) \ ,$$

where $z(\mathbf{s}_d) = Z((\mathbf{0}, \mathbf{s}_d])$.

In summary we have the following:

Theorem 4.3.1

$$\frac{1}{n^{d/2}\sigma}\mathbf{e}'_{nA}(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\epsilon} \;\Rightarrow\; Z(A) - \underbrace{\int\cdots\int}_{A}\int_0^1\cdots\int_0^1 g(\mathbf{s}_d,\mathbf{t}_d)dZ(\mathbf{t}_d)d\mathbf{s}_d \;,$$

where

$$g(\mathbf{s}_d,\mathbf{t}_d) = \mathbf{f}'(\mathbf{s}_d)\mathbf{G}\mathbf{f}(\mathbf{t}_d) \;.$$

We denote this limit process by $Z_g(A)$. That is,

$$Z_g(A) = Z(A) - \underbrace{\int\cdots\int}_{A}\int_0^1\cdots\int_0^1 g(\mathbf{s}_d,\mathbf{t}_d)dZ(\mathbf{t}_d)$$

The weak convergence results of Alexander and Pyke (1986) and the approach of MacNeill (1978 b), which applies Theorem 5.5 of Billingsley (1968), may be used to prove this result.

It may be shown that

$$E[Z_g(A)] = Z_g(\mathbf{0}) = 0 \;, \quad A \in \mathcal{A}$$

and the covariance kernel for any $A, B \in \mathcal{A}$ is

$$\begin{aligned} K(A,B) &= E[Z_g(A)Z_g(B)] \\ &= |A\cap B| - \underbrace{\int\cdots\int}_{A}\underbrace{\int\cdots\int}_{B} g(\mathbf{s}_d,\mathbf{t}_d)d\mathbf{s}_d d\mathbf{t}_d. \end{aligned}$$

In the null case, the regressor functions are equal to zero, then $Y_j = \epsilon_j$, and we have $Z_g(A) = Z(A)$, where Z is a Brownian process.

If the only regressor function is a constant, then $Y_j = \mu + \epsilon_j$, and we have

$$Z_g(A) = Z_0(A) = Z(A) - |A|Z(I^d) \;,$$

where Z_0 is a Brownian bridge process.

If we let $p(\omega_A)$ be the uniform distribution, then a test statistic for detecting change of intercept (i.e. β_0) in regression models indexed by A is defined as follows:

$$Q_{g_n} = \sum_{A} \left(\sum_{j \in nA} r_j \right)^2 ,$$

where r_j a is regression residual.

Under H_0,

$$\frac{1}{n^{2d}\sigma^2} U_g \rightarrow \underbrace{\int_0^1 \cdots \int_0^1}_{d} Z_g^2(A)dA = \int_{I^d} Z_g^2(A)dA , \quad A \in \mathcal{A}^* .$$

Large sample quantiles for Cramér von-Mises stochastic are obtained through evaluation of the distribution of relevant stochastic integrals. Anderson and Darling (1952), MacNeill (1974, 1978), Jandhyala and MacNeill(1989, 1991) and Jandhyala and Minogue (1993) developed several approaches for computing quantiles for the Cramér von-Mises stochastic integrals that arise in the large sample theory for change-point statistics. However, there are difficulties, mainly involving multiplicities of eigenvalues of the Karhünen-Loeve expansion of the limit process, in extending these approaches to higher dimensions.

In order to evaluate

$$P\left(\int_{I^d} Z_g^2(A)dA \leq q \right)$$

for given $q > 0$, and conversely q for given probability level α, we consider the method introduced by Imhof (1961) and used by Eastwood (1993) to obtain approximate

distributions as follows:

$$P\left(\int_{I^d} Z_g^2(A)dA \leq q\right) \simeq P\left(\chi_h^2 \leq q'\right) ,$$

where χ_h^2 denotes a central chi-square random variable with h degrees of freedom, where q and q' are related by the equation $q' = (q - c_1)\sqrt{\frac{h}{c_2}} + h$, where $h = \frac{c_2^3}{c_3^2}$ and where c_i $(i = 1, 2, 3)$ is the i^{th} cumulant.

Example 1: A Cramér-von Mises type statistic for detecting change of intercept in a regression model for two-dimensional data and boundaries as in Figure 2.1 is

$$Q_{gn} = \sum_{l=1}^{n}\sum_{k=1}^{n}\left(\sum_{i=1}^{l}\sum_{j=1}^{k} r_{ij}\right)^2 .$$

This statistic was considered by MacNeill and Jandhyala (1993) but was not derived for any particular set of boundaries nor was the large sample distribution obtained.

Under H_0,

$$\frac{1}{n^4\sigma^2}Q_{gn} \rightarrow \int_0^1\int_0^1 Z_g^2(t_1, t_2)dt_1dt_2 .$$

We apply the Imhof (1961) method to obtain quantiles for change-boundary statistics for the following models: When the model is a null, then

$$Y_{ij} = \epsilon_{ij} .$$

When the model is a constant, then

$$Y_{ij} = \beta_0 + \epsilon_{ij} .$$

When the model is linear, then

$$Y_{ij} = \beta_0 + \beta_1(i/n) + \beta_1(j/n) + \epsilon_{ij} .$$

When the model is a quadratic, then

$$Y_{ij} = \beta_0 + \beta_1(i/n) + \beta_1(j/n) + \beta_3(i/n)^2 + \beta_4(j/n)^2 + \beta(ij/n^2) + \epsilon_{ij} \ .$$

Table 1 contains selected quantiles for the distribution of

$$\int_0^1 \int_0^1 Z_g^2(t_1, t_2) dt_1 dt_2 \ ,$$

which is used for detecting a change of the intercept in regression models.

Table 1: Selected quantiles for testing parameter β_0.

$$P\left(\int_0^1 \int_0^1 Z_g^2(t_1, t_2) dt_1 dt_2 \leq q\right) = \alpha$$

α	null	constant	linear	quadratic
.005	.07640	.05897	.03206	.04381
.01	.07644	.05905	.03318	.04381
.025	.07668	.05943	.03542	.04381
.05	.07742	.06031	.03809	.04382
.1	.08011	.06270	.04195	.04382
.5	.16101	.10547	.06608	.04383
.9	.53805	.25966	.11053	.04560
.95	.72413	.33111	.12742	.05435
.975	.91678	.40397	.14371	.07246
.99	1.17825	.50169	.16461	.10799
.995	1.37971	.57639	.18007	.14101

If we let $p(\omega_A)$ be a uniformly distribution function, then the test statistic for detecting changes of all parameters in regression models given by

$$U_n = \sum_A \mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X}_A\mathbf{X}_A'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \quad ,$$

where $\mathbf{X}_A = (f_0(n^{-1}\mathbf{j}), \ldots, f_{p-1}(n^{-1}\mathbf{j}))$ with components replaced by 0 when $n^{-1}\mathbf{j}$ is not in A, can be written as follows:

$$U_n = \sum_A \left(\sum_{\mathbf{j} \in nA} f_0(n^{-1}\mathbf{j})r_\mathbf{j} \right)^2 + \cdots + \sum_A \left(\sum_{\mathbf{j} \in nA} f_{p-1}(n^{-1}\mathbf{j})r_\mathbf{j} \right)^2 \quad .$$

Let e_{nA}^h denote the n^d-dimensional stacked vector that has components equal to $h(n^{-1}\mathbf{j})$ if $n^{-1}\mathbf{j} \in A$ and equal to 0 otherwise. We can obtain similar results as follows:

Theorem 4.3.1 *Let the regression functions*

$$f_i(\mathbf{t}_d), \quad \mathbf{t}_d \in I_d, \quad i = 0, 1, \ldots, p-1 \ ,$$

and the function $h(\mathbf{t}_d)$ be continuously differentiable on I_d. Then the sequence of stochastic processes

$$\left\{ \frac{1}{n^{d/2}\sigma} S_{gn}^h(A) = \frac{1}{n^{d/2}\sigma} \mathbf{e}_{nA}^{h\prime}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\epsilon} \right\}$$

converges weakly to the Gaussian process $\{Z_g^h(A), \quad A \in \mathcal{A}^*\}$ defined by

$$Z_g^h(A) = \underbrace{\int \cdots \int}_{A} h(\mathbf{t}_d)dZ(\mathbf{t}_d) - \underbrace{\int \cdots \int}_{A} h(\mathbf{t}_d) \int_0^1 \cdots \int_0^1 g(\mathbf{s}_d, \mathbf{t}_d)dZ(\mathbf{t}_d)d\mathbf{s}_d \ .$$

Proof. The partial sums of residuals can be denoted as follows:

$$S_{gn}^h(A) = \mathbf{e}_{nA}^{h\prime}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\epsilon} \ .$$

Then

$$\mathbf{e}'_{nA}\mathbf{I}\epsilon = \sum_{\mathbf{j}\in nA} h(n^{-1}\mathbf{j})\epsilon_{\mathbf{j}} .$$

If the components of ϵ are iid with mean zero and finite variance σ^2, then it follows from the discussion in Section 4.2 that:

$$\frac{1}{n^{d/2}\sigma}\mathbf{e}^{h}_{nA}{}'\mathbf{I}\epsilon \Rightarrow \underbrace{\int\cdots\int}_{A} h(\mathbf{t}_d)dZ(\mathbf{t}_d) ,$$

Furthermore, it can be seen that

$$\frac{1}{n^{d/2}}\mathbf{e}^{h}_{nA}{}'\mathbf{X} \rightarrow \underbrace{\int\cdots\int}_{A} h(\mathbf{s}_d)\mathbf{f}(\mathbf{s}_d)d\mathbf{s}_d ,$$

Also, if $z(\mathbf{s}_d) = Z((\mathbf{0}, \mathbf{s}_d])$, then it can be seen that

$$\frac{1}{n^{d/2}\sigma}\mathbf{X}'\epsilon \Rightarrow \int_0^1\cdots\int_0^1 \mathbf{f}(\mathbf{s}_d)dZ(\mathbf{s}_d) .$$

If we use the same method in Theorem 4.3.1, we obtain

$$\frac{1}{n^{d/2}\sigma}S^{h}_{gn}(A) \Rightarrow Z^{h}_{g}(A) , \quad A \in \mathcal{A}^* .$$

Further the covariance kernel of the limit process is given by

$$\begin{aligned} K(A,B) &= E[Z^{h}_{g}(A)Z^{h}_{g}(B)] \\ &= \underbrace{\int\cdots\int}_{|A\cap B|} h^2(\mathbf{t}_d)d\mathbf{t}_d - \underbrace{\int\cdots\int}_{A} h(\mathbf{t}_d)\underbrace{\int\cdots\int}_{B} h(\mathbf{s}_d)g(\mathbf{s}_d,\mathbf{t}_d)d\mathbf{s}_d d\mathbf{t}_d. \end{aligned}$$

The test statistic for detecting change in only one coefficient in a regression model is

$$Q^{f_i}_{gn} = \sum_{A}\left(\sum_{\mathbf{j}\in nA} f_i(n^{-1}\mathbf{j})r_{\mathbf{j}}\right)^2 , \quad i = 0, 1, \ldots, p-1 ,$$

where $r_{\mathbf{j}}$ is the regression residual at $\mathbf{j}$.

Under H_0,

$$\frac{1}{n^{2d}\sigma^2}Q_{gn}^{f_i} \to \underbrace{\int_0^1 \cdots \int_0^1}_{d} \{Z_g^{f_i}(A)\}^2 dA = \int_{I^d} \{Z_g^{f_i}(A)\}^2 dA\,,\ i = 0, 1, \ldots, p-1\,,\ A \in \mathcal{A}^*\,.$$

The test statistic for detecting changes in all coefficients in regression models is

$$U_n = \sum_A \left(\sum_{\mathbf{j} \in nA} f_0(n^{-1}\mathbf{j}) r_{\mathbf{j}} \right)^2 + \cdots + \sum_A \left(\sum_{\mathbf{j} \in nA} f_{p-1}(n^{-1}\mathbf{j}) r_{\mathbf{j}} \right)^2 \quad .$$

It can be shown that, under H_0,

$$\frac{1}{n^{2d}\sigma^2}U_n \to \int_{I^d} \{Z_g^{f_0}(A)\}^2 dA \ + \ \cdots \ + \ \int_{I^d} \{Z_g^{f_{p-1}}(A)\}^2 dA\,, A \in \mathcal{A}^*\,.$$

Example 2: For detecting a change in either parameter β_1 or β_2 in regression models for two-dimensional data and boundaries as in Figure 2.1, the test statistics are

$$Q_{gn}^{f_u} = \sum_{l=1}^{n} \sum_{k=1}^{n} \left(\sum_{i=1}^{l} \sum_{j=1}^{k} f_u(i/n, j/n) r_{ij} \right)^2 \quad , \quad u = 1, 2, 3, 4, 5\,,$$

where $f_1(i/n, j/n) = \frac{i}{n}$, $f_2(i/n, j/n) = \frac{j}{n}$, $f_3(i/n, j/n) = (\frac{i}{n})^2$, $f_4(i/n, j/n) = (\frac{j}{n})^2$ and $f_5(i/n, j/n) = \frac{ij}{n^2}$.

Under H_0,

$$\frac{1}{n^4\sigma^2}Q_{gn}^{f_u} \to \int_0^1 \int_0^1 \{Z_g^{f_u}(t_1, t_2)\}^2 dt_1 dt_2\,.$$

Table 2: Selected quantiles for testing parameter β_1 or β_2.

$$P\left(\int_0^1 \int_0^1 \{Z_g^{f_u}(t_1,t_2)\}^2 dt_1 dt_2 \leq q\right) = \alpha,\ u = 1,2$$

α	linear	quadratic
.005	.00987	.00652
.01	.00994	.06690
.025	.01014	.00704
.05	.01046	.00745
.1	.01110	.00809
.5	.01735	.01216
.9	.03346	.01989
.95	.04029	.02286
.975	.04708	.02573
.99	.05602	.02942
.995	.06277	.03216

Table 3: Selected quantiles for testing parameter β_3 or β_4.

$$P\left(\int_0^1 \int_0^1 \{Z_g^{f_u}(t_1,t_2)\}^2 dt_1 dt_2 \leq q\right) = \alpha,\ u = 3,4$$

α	quadratic
.005	.00339
.01	.00344
.025	.00355
.05	.00370
.1	.00396
.5	.00609
.9	.01088
.95	.01283
.975	.01475
.99	.01724
.995	.01912

Table 4: Selected quantiles for testing parameter β_5.

$$P\left(\int_0^1 \int_0^1 \{Z_g^{f_5}(t_1,t_2)\}^2 dt_1 dt_2 \leq q\right) = \alpha,$$

α	quadratic
.005	.00154
.01	.00159
.025	.00168
.05	.00180
.1	.00196
.5	.00296
.9	.00477
.95	.00545
.975	.00611
.99	.00695
.995	.07570

Example 3 : For detecting changes in all parameters of a regression models for two-dimensional data and boundaries as in Figure 2.1, the test statistics are as follows:

When the model is linear, then

$$U_n^l = \sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{u=0}^{2}\left(\sum_{i=1}^{l}\sum_{j=1}^{k} f_u(i/n, j/n)r_{ij}\right)^2 .$$

When the model is quadratic, then

$$U_n^q = \sum_{l=1}^{n}\sum_{k=1}^{n}\sum_{u=0}^{4}\left(\sum_{i=1}^{l}\sum_{j=1}^{k} f_u(i/n, j/n)r_{ij}\right)^2 .$$

Under H_0,

$$\frac{1}{n^4\sigma^2}U_n^l \to \sum_{u=0}^{2}\int_0^1\int_0^1 \{Z_g^{f_u}(t_1,t_2)\}^2 dt_1 dt_2 ,$$

$$\frac{1}{n^4\sigma^2}U_n^q \to \sum_{u=0}^{4}\int_0^1\int_0^1 \{Z_g^{f_u}(t_1,t_2)\}^2 dt_1 dt_2 .$$

Table 5: Selected quantiles for testing changes of all parameters.

α	linear	quadratic
.005	.05195	.04148
.01	.05373	.04375
.025	.05729	.04783
.05	.06144	.05213
.1	.06773	.05809
.5	.10649	.08918
.9	.17811	.13890
.95	.20536	.15680
.975	.23165	.17378
.99	.26539	.19522
.995	.29036	.21091

4.4 Conclusion

The test statistics derived in Chapter 2 and Chapter 3 can be defined in terms of set indexed partial sums of regression residuals. Limit processes are developed for set indexed partial sums of regression residuals and large sample distributions of change-boundary statistics are those of certain functionals on Brownian processes indexed by $\mathcal{A}^*$. We compute and tabulate the distributions for those statistics defined by selected two-dimensional change-boundaries. Theory is developed to obtain the asymptotic distributions of the test statistics for general change-boundary (convex) in higher dimensions.

Chapter 5

LIMITS FOR THE RESIDUAL PROCESS OF STATIONARY SPATIAL SERIES

5.1 Introduction

Limit distributions of partial sum processes for spatial data have been computed under the assumption that the observations are independent. Tang and MacNeill (1993) discussed the limits of residual processes defined by the partial sums for stationary time series. MacNeill (1996) derived the properties of partial sums of squared residuals for non-normal and general serially correlated observations. However these papers involve only partial sum processes for the time series case.

In this chapter we extend the results of Tang and MacNeill (1993) to the spatial case. We consider linear regression of a random variable against general non-stochastic functions of a matrix array, but with error variables that form a stationary spatial process. We then examine the large sample properties of the stochastic process defined by the matrix array of the partial sums of regression residuals. After introducing the problem in section 5.2 we derive, in section 5.3, the residual processes for stationary spatial series satisfying a moment condition. These processes are used in section 5.4 to

obtain the residual processes for regression against general nonstochastic regression functions of a matrix array when the errors form a stationary spatial series. We then discuss in section 5.5 the effect of spatial serial correlation on change boundary statistics and large sample adjustments to account for this spatial serial correlation. For convenience we discuss only 2-dimensional spatial series.

5.2 Regression Models and Error Process Structure

We first define the basic model. Let $X(n,m)$ $(n,m = 0,\pm 1,\ldots)$ be a zero mean, stationary spatial series defined on a lattice with covariance function

$$R(u,v) = E\{X(t,s)X(t+u,s+v)\}, \quad |u|,\ |v| < \infty \ .$$

If the covariance function is absolutely summable, i.e.,

$$\sum_{u=-\infty}^{\infty}\sum_{v=-\infty}^{\infty} |R(u,v)| < \infty \ , \tag{5.1}$$

then the spectral density function,

$$f(\lambda_1,\lambda_2) = \frac{1}{4\pi^2}\sum_{|u|<\infty}\sum_{|v|<\infty} e^{-i\lambda_1 u - i\lambda_2 v}R(u,v) \ , \quad \lambda_1,\lambda_2 \in [-\pi,\pi] \ ,$$

exists.

In the sequel we require a central limit theorem for spatially correlated series. Brillinger (1970) defined the cumulant functions for stationary spatial series as follows:

$$C_{k+1}(\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_k) = Cum\{X(\mathbf{n}+\mathbf{v}_1),\ldots,X(\mathbf{n}+\mathbf{v}_k),X(\mathbf{n})\} \ ,$$

where $\mathbf{v}_j = (v_{1j},v_{2j})$, $\mathbf{n} = (n_1,n_2)$. Stationarity to order $k+1$ is implicit in this definition. Note that the first two cumulants are $E\{X(n_1,n_2)\}$ and $R(v_1,v_2)$, $|v_1|,|v_2| < \infty$.

When necessary we assume the cumulants exist and satisfy what we call the Brillinger condition for spatial data namely,

$$|C_{k+1}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)| < \frac{L_k}{\prod_{j=1}^{k}(1+v_{1j}^2)(1+v_{2j}^2)} \ , \tag{5.2}$$

for some finite L_k, where $\mathbf{v}_j = (v_{1j}, v_{2j})$, $j = 1, 2, \ldots, k$.

If (5.1) and (5.2) are satisfied, the results of Brillinger (1973) become as follows:

For $t, s \in [0, 1]$,

$$\frac{1}{n}\sum_{i=1}^{[nt]}\sum_{j=1}^{[ns]} X(i,j) \ ,$$

converges in distribution to the normal with zero mean and variance $\{4\pi^2 f(0,0)ts\}$.

We now consider the regression model

$$Y_n(i,j) = \sum_{k=0}^{p} \beta_k g_k(i/n, j/n) + X(i,j) \ ,$$

where $\{g_k(\cdot,\cdot),\ 0 \le k \le p\}$ is a collection of regressor functions defined on the unit square.

If we denote the vector of regression coeffients by $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$, the design matrix by $\mathbf{A}_n$, the stacked vector of observations by $\mathbf{Y}_n$ and the stacked vector of stationary spatial series by $\mathbf{X}_n$, then the model may be written in matrix form as

$$\mathbf{Y}_n = \mathbf{A}_n\boldsymbol{\beta} + \mathbf{X}_n \ .$$

The regression parameter estimators are

$$\hat{\boldsymbol{\beta}} = (\mathbf{A}_n'\mathbf{A}_n)^{-1}\mathbf{A}_n'\mathbf{Y}_n \ .$$

The matrix array of partial sums of regression residuals are defined as

$$S_{g_n}(k,l) = \sum_{i=1}^{k}\sum_{j=1}^{l}\{Y_n(i,j) - \hat{Y}_n(i,j)\} \ , \quad 1 \le k, l \le n,$$

where $\hat{Y}(i,j) = \boldsymbol{\beta}'\mathbf{g}(i/n, j/n)$ and $\mathbf{g}(i/n, j/n)' = (g_0(i/n, j/n), \ldots, g_p(i/n, j/n))$.

Since we shall be concerned with weak convergence on the space of functions continuous on the unit square, $C[0,1]^2$, we use these matrix arrays of partial sums to define a sequence of stochastic processes $\{Z_{g_n}(t,s),\ t,s \in [0,1]\}$ $(n \geq 1)$ possessing continuous sample paths as follows (see Kuelbs (1968)):

$$\begin{aligned} nZ_{g_n}(t,s) \;=\; & S_{g_n}([nt],[ns]) + (nt - [nt])\{S_{g_n}([nt]+1,[ns]) - S_{g_n}([nt],[ns])\} \\ & + (ns - [ns])\{S_{g_n}([nt],[ns]+1) - S_{g_n}([nt],[ns])\} \\ & + n(nt - [nt])(ns - [ns])\{Y_n([nt]+1,[ns]+1) \\ & - \hat{Y}_n([nt]+1,[ns]+1)\} \ . \end{aligned}$$

If we let $\mathbf{e}_{nt,ns}$ denote the n^2-dimensional vector that has: 1 for components where $\mathbf{X}_n$ has as its component $X_n(i,j)$ with $i \leq [nt]$ and $j \leq [ns]$, $nt - [nt]$ with $i = [nt]+1$ and $j \leq [ns]$, $ns - [ns]$ with $i \leq [nt]$ and $j = [ns]+1$, $n(nt - [nt])(ns - [ns])$ with $i = [nt]+1$ and $j = [ns]+1$, and 0 otherwise, then we can write

$$nZ_{g_n}(t,s) = \mathbf{e}'_{nt,ns}\{\mathbf{I} - \mathbf{A}_n(\mathbf{A}'_n\mathbf{A}_n)^{-1}\mathbf{A}'_n\}\mathbf{X}_n \ .$$

5.3 The Partial Sum Limit Process for Stationary Spatial Series

To establish the limit process for $\{Z_{g_n}(t,s),\ t,s \in [0,1]\}$ we need first to examine the properties of the matrix array of partial sums of the error process $X(n,m)$ $(n,m = 0, \pm 1, \ldots)$ Hence, we let $S_{X_n}(k,l) = \sum_{i=1}^{k}\sum_{j=1}^{l} X(i,j)$ and define another sequence of stochastic processes $\{Z_{X_n}(t,s),\ t,s \in [0,1]\}$ $(n \geq 1)$ possessing continuous sample

paths by

$$\begin{aligned} nZ_{X_n}(t,s) &= S_{X_n}([nt],[ns]) + (nt-[nt])\{S_{X_n}([nt]+1,[ns]) - S_{X_n}([nt],[ns])\} \\ &\quad + (ns-[ns])\{S_{X_n}([nt],[ns]+1) - S_{X_n}([nt],[ns])\} \\ &\quad + n(nt-[nt])(ns-[ns])X_n([nt]+1,[ns]+1) \ . \end{aligned}$$

We note first that

$$Z_{X_n}(0,0) = E\{Z_{X_n}(t,s)\} = 0$$

and consider next the covariance kernel of the process

$$K_n(t_1,s_1;t_2,s_2) = E\{Z_{X_n}(t_1,s_1)Z_{X_n}(t_2,s_2)\} \ .$$

We assume $t_1 = \min(t_1,t_2)$, $s_1 = \min(s_1,s_2)$, $k_i = [nt_i]$ and $l_i = [ns_i]$, $i = 1,2$.

For sufficient large n, we have

$$\left| K_n(t_1,s_1;t_2,s_2) - E\left\{ Z_{X_n}(\frac{k_1}{n},\frac{l_1}{n}) Z_{X_n}(\frac{k_2}{n},\frac{l_2}{n}) \right\} \right| \le \frac{c}{n} \ ,$$

where $c > 0$ is independent of t_1, s_1, t_2, s_2 and n.

Therefore, for large samples, we need only consider $K_n\left(\frac{k_1}{n},\frac{l_1}{n};\frac{k_2}{n},\frac{l_2}{n}\right)$. Then

$$\begin{aligned} K_n\left(\frac{k_1}{n},\frac{l_1}{n};\frac{k_2}{n},\frac{l_2}{n}\right) &= \frac{1}{n^2} E\{S_{X_n}(k_1,l_1)S_{X_n}(k_2,l_2)\} \\ &= \frac{1}{n^2} \sum_{t_1=1}^{l_1} \sum_{s_1=1}^{k_1} \sum_{t_2=1}^{l_2} \sum_{s_2=1}^{k_2} E\{X(t_1,s_1)X(t_2,s_2)\} \\ &= \frac{1}{n^2} \sum_{t_1=1}^{l_1} \sum_{s_1=1}^{k_1} \sum_{t_2=1}^{l_2} \sum_{s_2=1}^{k_2} R(t_2-t_1,s_2-s_1) \ . \end{aligned}$$

Conditions (5.1) and (5.2) imply

$$\begin{aligned} &\frac{1}{n^2} \sum_{t_1=1}^{l_1} \sum_{t_2=1}^{l_2} \sum_{s_1=1}^{k_1} \sum_{s_2=1}^{k_2} R(t_2-t_1,s_2-s_1) \\ &\qquad = \frac{1}{n^2} \sum_{|u|<l_1} \sum_{|v|<k_1} (l_1-|u|)(k_1-|v|)R(u,v) + O(n^{-1}) \ . \end{aligned}$$

We have shown that

$$\frac{1}{n^2}K_n(t_1,s_1;t_2,s_2) \to t_1 s_1 \sum_{u=-\infty}^{\infty}\sum_{v=-\infty}^{\infty} R(u,v) \ ,$$

where $t_1 = \min(t_1,t_2)$ and $s_1 = \min(s_1,s_2)$.

That is,

$$\frac{1}{n^2}K_n(t_1,s_1;t_2,s_2) \to 4\pi^2 f(0,0)(t_1 \wedge t_2)(s_1 \wedge s_2) \ ,$$

where $t_1 \wedge t_2 = \min(t_1,t_2)$.

We adapt the method of MacNeill (1996) to establish asymptotic normality in the following theorem.

Theorem 5.3.1 *Under assumptions (5.1) and (5.2), the p-vector* $\{Z_{X_n}(t_1,s_1),\ldots,$ $Z_{X_n}(t_p,s_p)\}$ *has a non-trivial asymptotic probability distribution that is normal with zero mean and covariance matrix* $\|4\pi^2 f(0,0)(t_i \wedge t_j)(s_i \wedge s_j)\|_{ij}$

Proof: The Brillinger condition for spatial data (5.2) can be used to demonstrate that the spatial cumulants of orders higher than two of a vector component of $Z_{X_n}(t_i,s_i)$ are $O(n^{-1})$ or smaller and hence that $Z_{X_n}(t_i,s_i)$ converges in distribution to the normal with zero mean and variance $4\pi^2 f(0,0)t_i s_i$. The Cramér-Wold device of demonstrating asymptotic multivariate normality by showing asymptotic normality with zero mean and variance

$$4\pi^2 f(0,0)\sum_{i=1}^{p}\sum_{j=1}^{p}\lambda_i\lambda_j(t_i \wedge t_j)(s_i \wedge s_j)$$

of $\sum_{i=1}^{p}\lambda_i Z_{X_n}(t_i,s_i)$, where the λ_i are arbitrary real numbers, can be used to complete the proof for the p-dimensional case.

We next show tightness of the sequence of measures P_{X_n} $(n = 1, 2, \ldots)$ generated in $C[0,1]^2$ by $\{Z_{X_n}(t,s),\ t,s \in [0,1]\}$. The arguments used above to derive the covariance kernel for these processes can be used to show the existence of a constant C such that for $t_1, s_1, t_2, s_2 \in [0,1]$,

$$E\{Z_{X_n}(t_1,s_1) - Z_{X_n}(t_2,s_2)\}^4 \leq C\{(t_1-t_2)^2 + (s_1-s_2)^2\} \tag{5.3}$$

where C is not dependent on t_1, s_1, t_2, s_2 and n.

We only discuss the case of $t_2 \geq t_1$ and $s_2 \geq s_1$ (the same argument holds for the other cases).

Let $[nt_1] = k_1$, $[ns_1] = l_1$, $[nt_2] = k_2$, $[ns_2] = l_2$, then

$$\begin{aligned}
& E\left\{\frac{1}{n}S_{X_n}(k_1,l_1) - \frac{1}{n}S_{X_n}(k_2,l_2)\right\}^4 \\
= \ & \frac{1}{n^4}E\left\{\left(\sum_{i=k_1+1}^{k_2}\sum_{j=l_1+1}^{l_2} + \sum_{i=1}^{k_1}\sum_{j=l_1+1}^{l_2} + \sum_{i=k_1+1}^{k_2}\sum_{j=1}^{l_1}\right) X_n(i,j)\right\}^4 \\
= \ & \frac{1}{n^4}\left(\sum_{A_1} + \sum_{A_2} + \sum_{A_3}\right) E\{X_n(i_1,j_1)X_n(i_2,j_2)X_n(i_3,j_3)X_n(i_4,j_4)\}\ ,
\end{aligned}$$

where

$$A_1 = \{(i_h,j_h):\ \ k_1 < i_h \leq k_2,\ l_1 < i_h \leq l_2,\ \ h = 1,2,3,4\}\ ,$$

$$A_2 = \{(i_h,j_h):\ \ 1 \leq i_h \leq k_1,\ l_1 < i_h \leq l_2,\ \ h = 1,2,3,4\}\ ,$$

$$A_3 = \{(i_h,j_h):\ \ k_1 < i_h \leq k_2,\ 1 \leq i_h \leq l_1,\ \ h = 1,2,3,4\}\ .$$

These fourth order moments can be expressed in terms of the corresponding fourth order cumulants and products of pairs of elements from the covariance function. Hence,

$$\frac{1}{n^4}\sum_{A_i} E\{X_n(i_1,j_1)X_n(i_2,j_2)X_n(i_3,j_3)X_n(i_4,j_4)\}$$

$$= \frac{1}{n^4}\sum_{A_i} Cum\{X_n(i_1,j_1), X_n(i_2,j_2), X_n(i_3,j_3), X_n(i_4,j_4)\}$$
$$+ \frac{1}{n^4}\sum_{A_i}\{R(i_1-i_2, j_1-j_2)R(i_3-i_4, j_3-j_4)$$
$$+ R(i_1-i_3, j_1-j_3)R(i_2-i_4, j_2-j_4)$$
$$+ R(i_1-i_4, j_1-j_4)R(i_2-i_3, j_2-j_3)\} ,$$

where A_i, $i = 1, 2, 3$.

Using the Brillinger condition for spatial data (5.2) and adapting the method given by Tang and MacNeill (1993), we can obtain

$$\frac{1}{n^4}\sum_{A_1} E\{X_n(i_1,j_1)X_n(i_2,j_2)X_n(i_3,j_3)X_n(i_4,j_4)\} \leq C_1'\left(\frac{k_1}{n}-\frac{k_2}{n}\right)^2\left(\frac{l_1}{n}-\frac{l_2}{n}\right)^2$$
$$\leq C_1\left\{\left(\frac{k_1}{n}-\frac{k_2}{n}\right)^2+\left(\frac{l_1}{n}-\frac{l_2}{n}\right)^2\right\} ,$$

$$\frac{1}{n^4}\sum_{A_2} E\{X_n(i_1,j_1)X_n(i_2,j_2)X_n(i_3,j_3)X_n(i_4,j_4)\} \leq C_2\left(\frac{l_1}{n}-\frac{l_2}{n}\right)^2 ,$$

$$\frac{1}{n^4}\sum_{A_3} E\{X_n(i_1,j_1)X_n(i_2,j_2)X_n(i_3,j_3)X_n(i_4,j_4)\} \leq C_3\left(\frac{k_1}{n}-\frac{k_2}{n}\right)^2 ,$$

where C_1, C_2 and C_3 are not dependent on t_1, s_1, t_2, s_2 and n.

Therefore, we can choose C independent of t_1, s_1, t_2, s_2 and n such that

$$E\left\{\frac{1}{n}S_{X_n}(k_1,i_1)-\frac{1}{n}S_{X_n}(k_2,l_2)\right\}^4 \leq C\left\{\left(\frac{k_1}{n}-\frac{k_2}{n}\right)^2+\left(\frac{l_1}{n}-\frac{l_2}{n}\right)^2\right\} .$$

If the process $\{Z_X(t,s),\ t,s \in [0,1]\}$ is defined by

$$Z_X(t,s) = \{4\pi^2 f(0,0)\}^{\frac{1}{2}} Z(t,s) ,$$

where $Z(t,s)$ is a Brownian sheet and if W_X is the measure in $C[0,1]^2$ corresponding to $Z_X(\cdot,\cdot)$, then we have the following result.

Theorem 5.3.2 *Under assumption (5.1) and (5.2),*

$$P_{X_n} \Longrightarrow W_x \ .$$

Proof: Theorem 5.3.1 assures us that the finite dimensional distributions of P_{X_n} converge to those of W_X, and (5.3) implies that the sequence P_{X_n} $(n = 1, 2, \ldots)$ is tight. The proof is completed by applying Theorem 12.3 of Billingsley (1968).

5.4 The Regression Residual Process for Stationary Spatial Error Structure

We now consider the matrix array of partial sums of regression residuals when the error process is a stationary spatial series. The vector of regressor functions evaluated at (t, s) is denoted by $\mathbf{f}(t, s) = (f_0(t, s), \ldots, f_p(t, s))'$. It may be seen that the matrix

$$\lim_{n\to\infty} \frac{1}{n^2}(\mathbf{A}_n'\mathbf{A}_n) \equiv G$$

has as its (i, j)th component

$$\int_0^1 \int_0^1 f_i(t, s) f_j(t, s) dt ds \ .$$

The inverse of G exists provided the regressor functions are linearly independent and square integrable; with this proviso, we define a multilinear form, $g(t_1, s_1; t_2, s_2)$, as follows:

$$g(t_1, s_1; t_2, s_2) = \mathbf{f}'(t_1, s_1) G^{-1} \mathbf{f}(t_2, s_2) \ .$$

Then we define a limit process $\{Z_{X_g}(t, s),\ t, s \in [0, 1]\}$ by

$$Z_{X_g}(t, s) = Z_X(t, s) - \int_0^t \int_0^s \int_0^1 \int_0^1 g(t_1, s_1; t_2, s_2) dZ_X(t_2, s_2) dt_1 ds_1 \ ,$$

where $Z_X(t,s) = \sqrt{4\pi^2 f(0,0)}Z(t,s)$ and $Z(t,s)$ is the Brownian sheet.

The partial sum process of regression residuals is given by

$$nZ_{X_{g_n}}(t,s) = \mathbf{e}'_{nt,ns}\{\mathbf{I} - \mathbf{A}_n(\mathbf{A}'_n\mathbf{A}_n)^{-1}\mathbf{A}'_n\}\mathbf{X}_n \ .$$

If we apply arguments similar to those of Chapter 4, we obtain the following result.

Theorem 5.4.1 *Assume conditions (5.1) and (5.2). Further assume* $g_k(t,s)$ $(k = 0,1,\ldots,p)$ *are linearly independent non-stochastic regressor functions that are continuously differentiable on* $[0,1]^2$. *Then*

$$Z_{X_{g_n}}(t,s) \Longrightarrow Z_{X_g}(t,s) \ .$$

It can be shown that

$$E\{Z_{X_g}(t,s)\} = Z_{X_g}(0,0) = 0, \ t,s \in [0,1]$$

and that the covariance kernel for any $t_1, s_1, t_2, s_2 \in [0,1]$ is

$$\begin{aligned} K(t_1,s_1;t_2,s_2) &= E\{Z_g(t_1,s_1)Z_g(t_2,s_2)\} \\ &= 4\pi^2 f(0,0)\{(t_1 \wedge t_2)(s_1 \wedge s_2) \\ &\quad - \int_0^{t_1}\int_0^{s_1}\int_0^{t_2}\int_0^{s_2} g(t_1,s_1;t_2,s_2)dt_1 ds_1 dt_2 ds_2\} \ . \end{aligned}$$

5.5 Effect of Spatial Autocorrelation on Change Detection Statistics

In previous chapters we discussed tests for change boundaries for regression residuals when the noise is i.i.d. To deal with the problem of spatially correlated errors in regression, a model with stationary spatial error structure is considered. The effect of

autocorrelated errors on these statistics is discussed. We adapt the method of Tang and MacNeill (1993) to the spatial case.

We consider statistics derived in previous chapters for testing for parameter change. For the case of i.i.d error structure with $\sigma^2 < \infty$, a statistic for detecting change at unknown boundary as in Figure 2.1 in regression parameters is shown to be

$$Q_{g_n} = \frac{1}{n^4\sigma^2}\sum_{l=1}^{n}\sum_{k=1}^{n}\left\{\sum_{i=1}^{l}\sum_{j=1}^{k}[Y_n(i,j) - \hat{Y}_n(i,j)]\right\}^2 .$$

To make the statistic both operational and effective it is necessary to estimate σ^2 with an estimator that is consistent under both null and alternative hypotheses. Now assume the spatial error process is not i.i.d and $R(0,0)$ is used in place of σ^2. Note that

$$R(0,0) = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2 .$$

Then, if $Z_g(t,s) = \{4\pi^2 f(0,0)\}^{-1/2} Z_{X_g}(t,s)$,

$$Q_{g_n} \to \frac{4\pi^2 f(0,0)}{\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2}\int_0^1\int_0^1 Z_g^2(t,s)dtds .$$

The distribution of $\int_0^1\int_0^1 Z_g^2(t,s)dtds$ has been tabulated in Chapter 4. It indicates that the large sample effects of spatial serial correlation on Q_{g_n} can be adjusted for precisely by multiplying the quantiles of distributions for the i.i.d case by

$$\frac{4\pi^2 f(0,0)}{\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2} .$$

EXAMPLE:

If the noise process is a spatial unilateral multiplicative first-order autoregression

(Martin (1979)), that is,

$$X(t,s) + aX(t-1,s) + bX(t,s-1) + abX(t-1,s-1) = \epsilon(t,s) .$$

We have

$$f(\lambda_1,\lambda_2) = \frac{\sigma^2}{4\pi^2}(1 + 2a\cos\lambda_1 + a^2)^{-1}(1 + 2b\cos\lambda_2 + b^2)^{-1}$$

and

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2 = \frac{\sigma^2}{(1-a^2)(1-b^2)} .$$

Then

$$\frac{4\pi^2 f(0,0)}{\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2} = \frac{(1-a)(1-b)}{(1+a)(1+b)} .$$

5.6 Conclusion

Limit processes for the matrix array of partial sums of residuals in stationary spatial series and regression residuals with spatially correlated errors are developed.

Chapter 6

BOUNDARY ESTIMATION

6.1 Introduction

MacNeill (1993 a, b) and MacNeill and Jandhyala (1993) proposed ad hoc local procedures for estimating the locations of change-boundaries in spatial data. These methods reduce substantially the amount of computing required by estimation procedures that make few or no restrictions on the size of the collection of possible boundaries. Computing becomes an important issue as the grid size in the lattice shrinks. In Section 6.2 we review the nonparametric approach for boundary estimation. In Section 6.3 we discuss estimation of the location of boundaries by likelihood methods.

6.2 Nonparametric Approach

The problem of estimating the location of a change-boundary, given that one is present, was considered in a nonparametric setting by Carlstein and Krishnamoorthy (1992). Their approach uses empirical distribution functions for observations assumed to be from one distribution on one side of the boundary and from a different distribution on the other side. The selected boundary maximizes one of several proposed

criteria. We summarise these as follows:

Consider an array of independent random variables Y_j taken from the d-dimensional unit cube $I_d = [0,1]^d$. The unknown boundary B is a $(d-1)$-dimensional surface that partitions I_d into two regions, R_B and R_B. Assume Y_j are identically distributed with distribution function F when $j \in R_B$, and with G when $j \in R_B$ and $F \neq G$.

Let $\mathcal{B}$ be a finite collection of estimates for the unknown boundary B. For a candidate boundary B, compute the empirical cumulative distribution function of R_B

$$h_B = \frac{\sum_{j \in R_B} I(Y_j \leq x)}{|R_B|},$$

where $I(.)$ is an indicator function and $|R_B|$ is the number of observations in R_B, which treats all observations from region R_B as if they were identically distributed; similarly compute

$$h_B = \frac{\sum_{j \in R_B} I(Y_j \leq x)}{|R_B|},$$

which treats all observations from region R_B as if they were identically distributed.

Consider the differences

$$d_j^B = |h_B(Y_j) - h_B(Y_j)|$$

for each $j \in J$, where J is the collection of all observation indices. The boundary estimator B^* is the candidate boundary in $\mathcal{B}$ that maximizes the criterion function

$$D(B) = \left(\frac{|R_B|}{|J|}\right)\left(\frac{|R_B|}{|J|}\right) S(d_1^B, d_2^B, \ldots, d_{|J|}^B)$$

over all $B \in \mathcal{B}$, where $S(.)$ is a norming function which must satisfy certain simple

conditions Special cases include the Kolmogorov-Smirnov norm

$$S_{ks}(d_1, \ldots, d_{|J|}) = \sup_{1 \le i \le |J|} \{d_i\} ,$$

the Cramér-von Mises norm

$$S_{cs}(d_1, \ldots, d_{|J|}) = \left(\sum_{1 \le i \le |J|} \frac{d_i^2}{|J|} \right)^{\frac{1}{2}} ,$$

and the arithmetic-mean norm

$$S_{am}(d_1, \ldots, d_{|J|}) = \sum_{1 \le i \le |J|} \frac{d_i}{|J|} .$$

6.3 Likelihood Methods

Quandt (1958) and Worsley (1983) discussed maximum likelihood methods for estimating an unknown change-point. Esterby and El-Shaarawi (1981) proposed estimates of change-points based on maximizing marginal and conditional likelihood functions. In this section we extend these methods for estimating the location of change-points to estimating the location of boundaries in spatial data and we derive a statistic for boundary estimation.

6.3.1 Maximum likelihood method

Consider an array of n^d observations taken from the following model:

$$Y(\mathbf{j}) = \sum_{k=0}^{p-1} \beta_k x_k(\mathbf{j}) + \epsilon(\mathbf{j}) , \ \mathbf{j} \in R_B$$

and

$$Y(\mathbf{j}) = \sum_{k=0}^{p-1} \gamma_k x_k(\mathbf{j}) + \epsilon(\mathbf{j}) , \ \mathbf{j} \in R_B$$

where B is any unknown boundary, R_B is a region enclosed by B, $R_{\bar{B}}$ is the complementary set of R_B and $\epsilon(\mathrm{j}) \sim N(0, \sigma^2)$, $1 \leq \mathrm{j} \leq n1$.

The joint likelihood for β, γ, σ^2 and B based on the above model is

$$L = (2\pi\sigma^2)^{-\frac{n^*}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{\mathrm{j}\in R_B}(Y(\mathrm{j}) - \mathbf{X}_1(\mathrm{j})'\beta)^2 - \frac{1}{2\sigma^2}\sum_{\mathrm{j}\in R_{\bar{B}}}(Y(\mathrm{j}) - \mathbf{X}_2(\mathrm{j})'\gamma)^2\right\},$$

where $\beta = (\beta_0, \ldots, \beta_{p-1})'$, $\gamma = (\gamma_0, \ldots, \gamma_{q-1})'$, $\mathbf{X}_1(\mathrm{j}) = (x_0(\mathrm{j}), \ldots, x_{p-1}(\mathrm{j}) \ : \ \mathrm{j} \in R_B)'$ and $\mathbf{X}_2(\mathrm{j}) = (x_0(\mathrm{j}), \ldots, x_{q-1}(\mathrm{j}) \ : \ \mathrm{j} \in R_{\bar{B}})'$.

For a specified boundary B, β, γ and σ^2 can be separatedly estimated by the maximum likelihood method for the two regions R_B and $R_{\bar{B}}$, denoted by $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\sigma^2}$. If we replace the parameters β, γ and σ^2 by their estimates, $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\sigma^2}$, the likelihood function of B becomes

$$L(B) \propto \left(\sum_{\mathrm{j}\in R_B} r_{\mathrm{j}}^2 + \sum_{\mathrm{j}\in R_{\bar{B}}} r_{\mathrm{j}}^2\right)^{-\frac{n^*}{2}},$$

where r_{j} is the regression residual.

Hence the maximum likelihood estimator of the boundary B is that boundary B^* which minimizes the residual sum of squares, i.e.,

$$S^2_{R_{B^*}, R_{\bar{B}^*}} = \min_B (S^2_{R_B} + S^2_{R_{\bar{B}}})$$

where $S^2_{R_B} = \sum_{\mathrm{j}\in R_B} r_{\mathrm{j}}^2$ and $S^2_{R_{\bar{B}}} = \sum_{\mathrm{j}\in R_{\bar{B}}} r_{\mathrm{j}}^2$.

If σ^2 is known (set $\sigma^2 = 1$), the likelihood is

$$L(B) \propto \exp\left(-\frac{1}{2}\sum_{\mathrm{j}\in R_B} r_{\mathrm{j}}^2 - \frac{1}{2}\sum_{\mathrm{j}\in R_{\bar{B}}} r_{\mathrm{j}}^2\right).$$

The maximum likelihood estimator of a boundary corresponds to minimization of the residual sum of squares.

6.3.2 Marginal and Conditional Likelihood Methods

The marginal and conditional likelihood methods of Esterby and El-Shaarawi (1981) can be used to estimate the location of boundaries in spatial data as follows:

Given boundary B, the marginal and conditional likelihood are given by

$$L(B) \propto \left(\sum_{j \in R_B} r_j^2 + \sum_{j \in R_B} r_j^2 \right)^{-\frac{n^d - p - q - 2}{2}} .$$

Therefore, the maximum marginal and conditional likelihood estimator of a boundary again corresponds to the minimization of the residual sum of squares.

6.4 Conclusion

We discuss methods for estimating the locations of boundaries after detecting the presence of a change-boundary. Although these methods can theoretically be applied to quite general cases, there are still some difficulties for practical reasons. For example, for even moderated sized n, the number of possible boundaries to be considered becomes unmanageably large, thus making impractical direct application of these methods. The amount of computing required can be reduced by using the procedure proposed by MacNeill (1993a,b) and MacNeill and Jandhyala (1993).

Chapter 7

APPLICATIONS

We consider the wheat-yield data compiled by Mercer and Hall (1911) and discussed by Cressie (1993). The data are yields of grain (in pounds) for a 20×25 lattice of plots with 20 rows of plots running east to west and 25 columns of plots running north to south. Figure 7.1 presents the data and the fitted mean value function.

We test the data for change of mean level at unknown boundary. The collection of boundaries we consider first are those defined by the rectangles with sides parallel to those of the field and with one corner fixed in the north-west corner of the field. These are examples of the first set of rectangular boundaries discussed in Chapter 2.3 and illustrated in Figure 2.1.

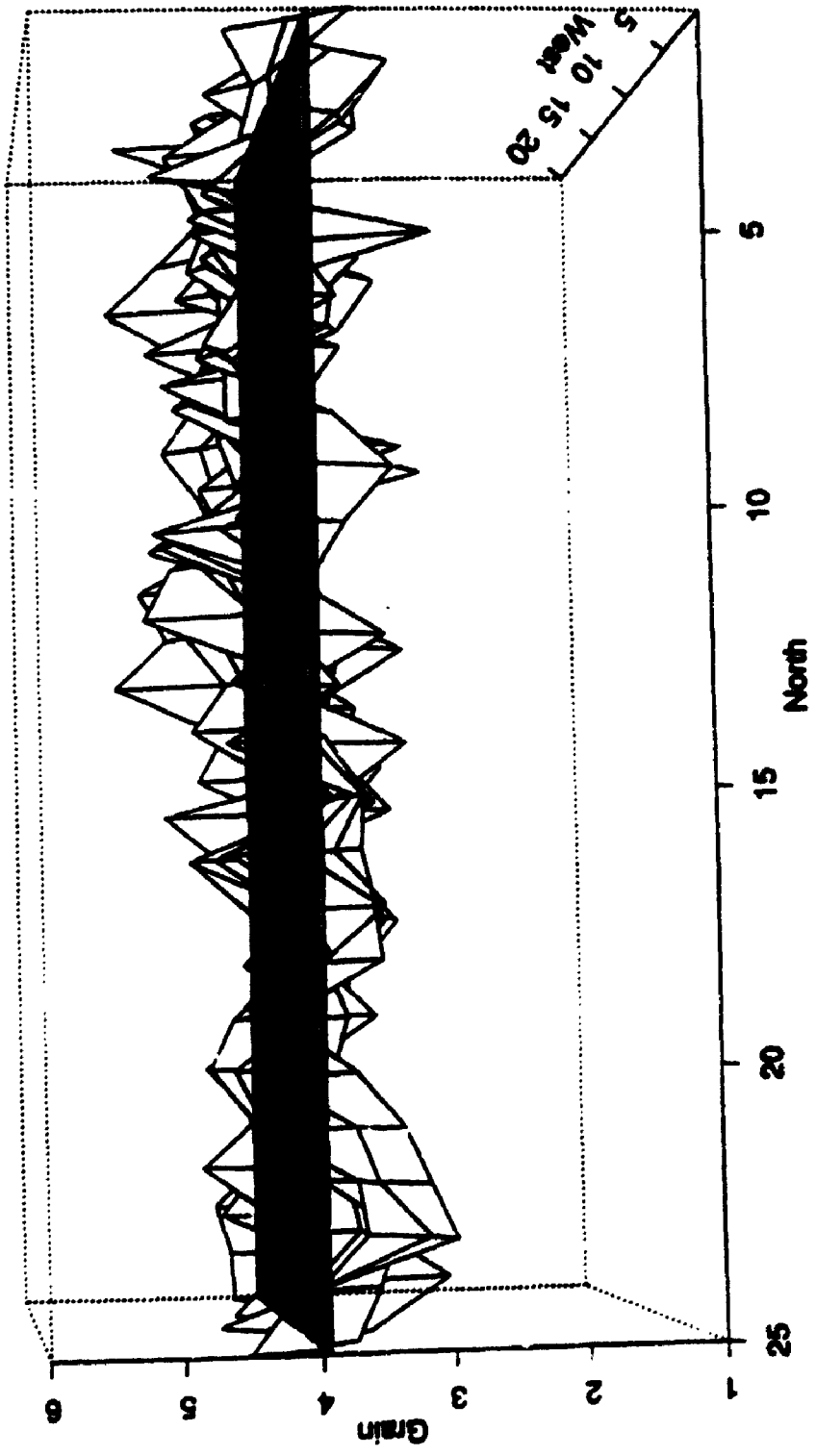

Figure 7.1: Mercer and Hall wheat-yield data and fitted model for no change in mean.

A locally optimal statistic for detecting change in mean for spatial data of this type is analogous to Q_{1n} in Chapter 2.3 and is defined as follows:

$$Q_{1cn} = \sum_{l=1}^{n} \sum_{k=1}^{[cn]} \left(\sum_{i=1}^{l} \sum_{j=1}^{k} (Y_{ij} - \bar{Y}) \right)^2 ,$$

where $c > 0$ is a constant and $[cn]$ is the largest integer in cn; this choice of lattice permits the use of statistics based on rectangular data sets which retain the same large sample properties as those of Chapter 2.

The hypotheses we are testing are as follows:

$$H_0 : \quad E(Y_{ij}) = \mu_0, \quad \text{for} \quad \text{all } (i,j)$$

versus

$$H_1 : \quad E(Y_{ij}) = \begin{cases} \mu_0 & \text{for} \quad \text{all } (i,j) \leq (l,k) \\ \mu_1 \neq \mu_0 & \text{otherwise.} \end{cases}$$

Under H_0, it can be shown that

$$\frac{1}{c^2 n^4 \sigma^2} Q_{1cn} \rightarrow \int_0^1 \int_0^1 Z_g^2(t_1, t_2) dt_1 dt_2 .$$

If σ^2 is not known, then an estimator that is consistent under both H_0 and H_1 can be obtained using differences in a regular grid. For example, if

$$d_{ij} = Y_{ij} - \frac{1}{4}(Y_{i-1,j} + Y_{i,j-1} + Y_{i+1,j} + Y_{i,j+1}) ,$$

then $E(d_{ij}^2) = \frac{5}{4}\sigma^2$. Hence we estimate σ^2 by $\tilde{\sigma}^2$ where

$$\tilde{\sigma}^2 = \frac{4}{5} \frac{1}{(n-2)([cn]-2)} \sum_{i=2}^{n-1} \sum_{j=2}^{[cn]-1} d_{ij}^2 .$$

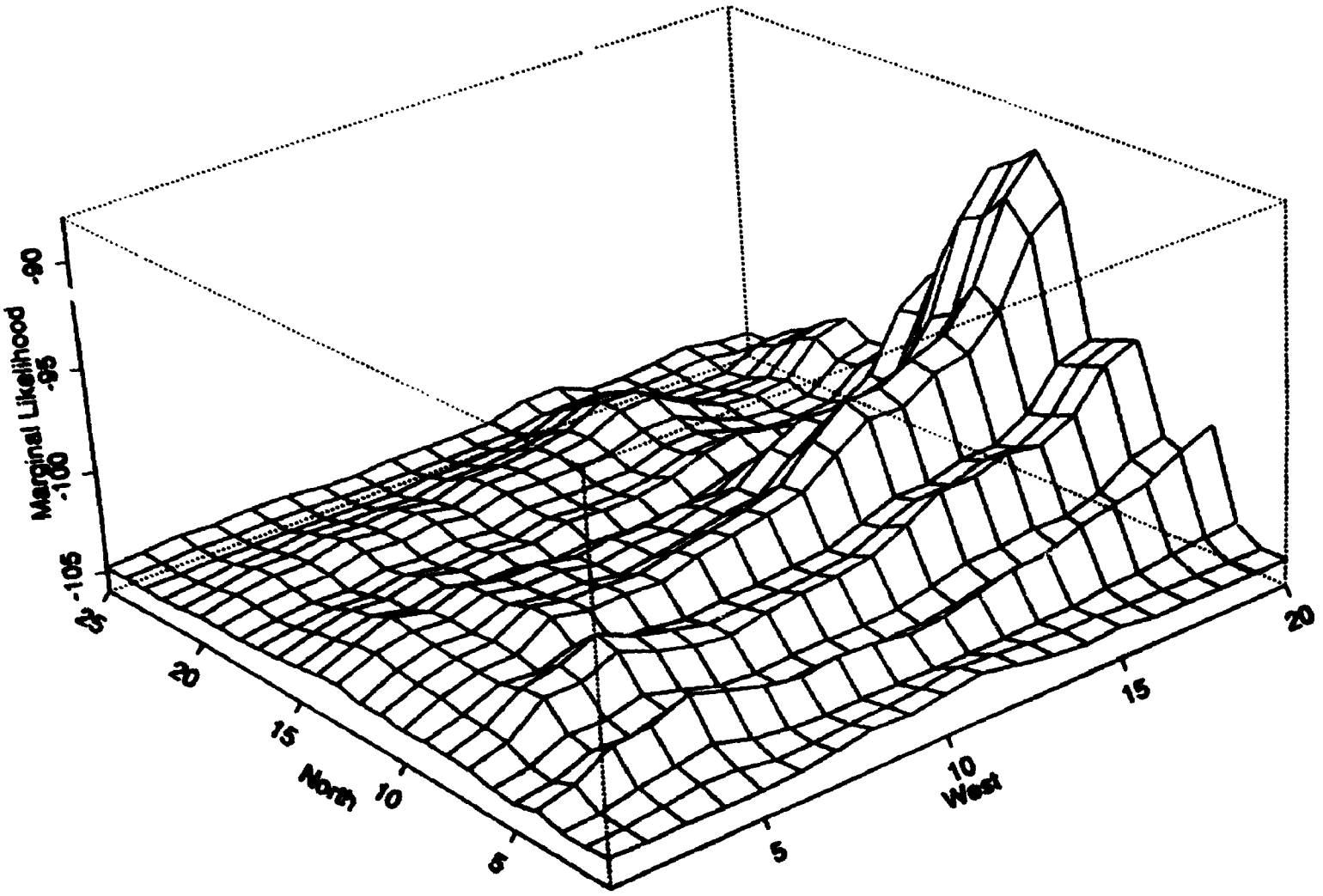

Figure 7.2: Marginal likelihood for the boundary location in the Mercer and Hall data.

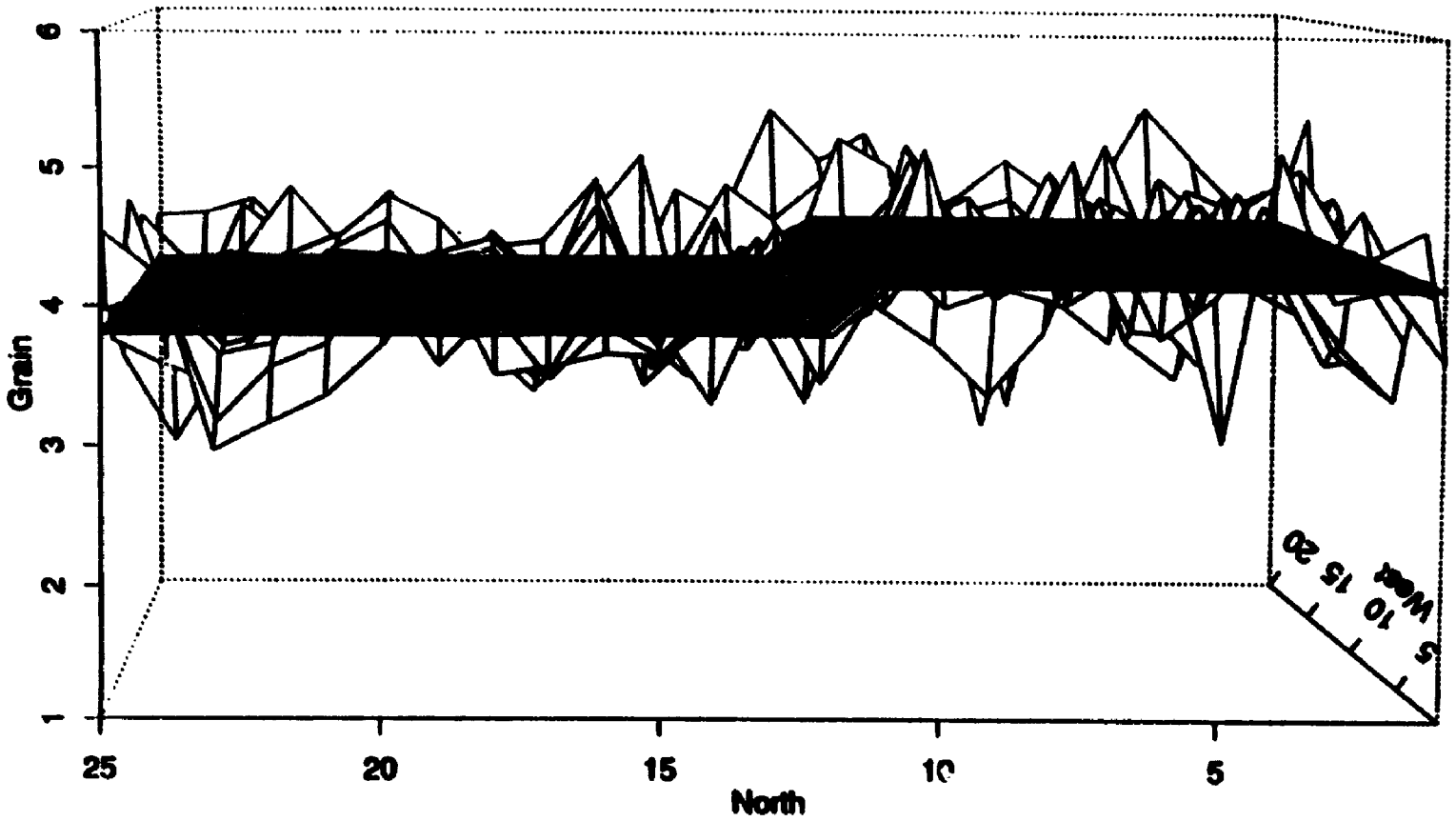

Figure 7.3: Mercer and Hall wheat-yield data and fitted change-boundary model based on B^*.

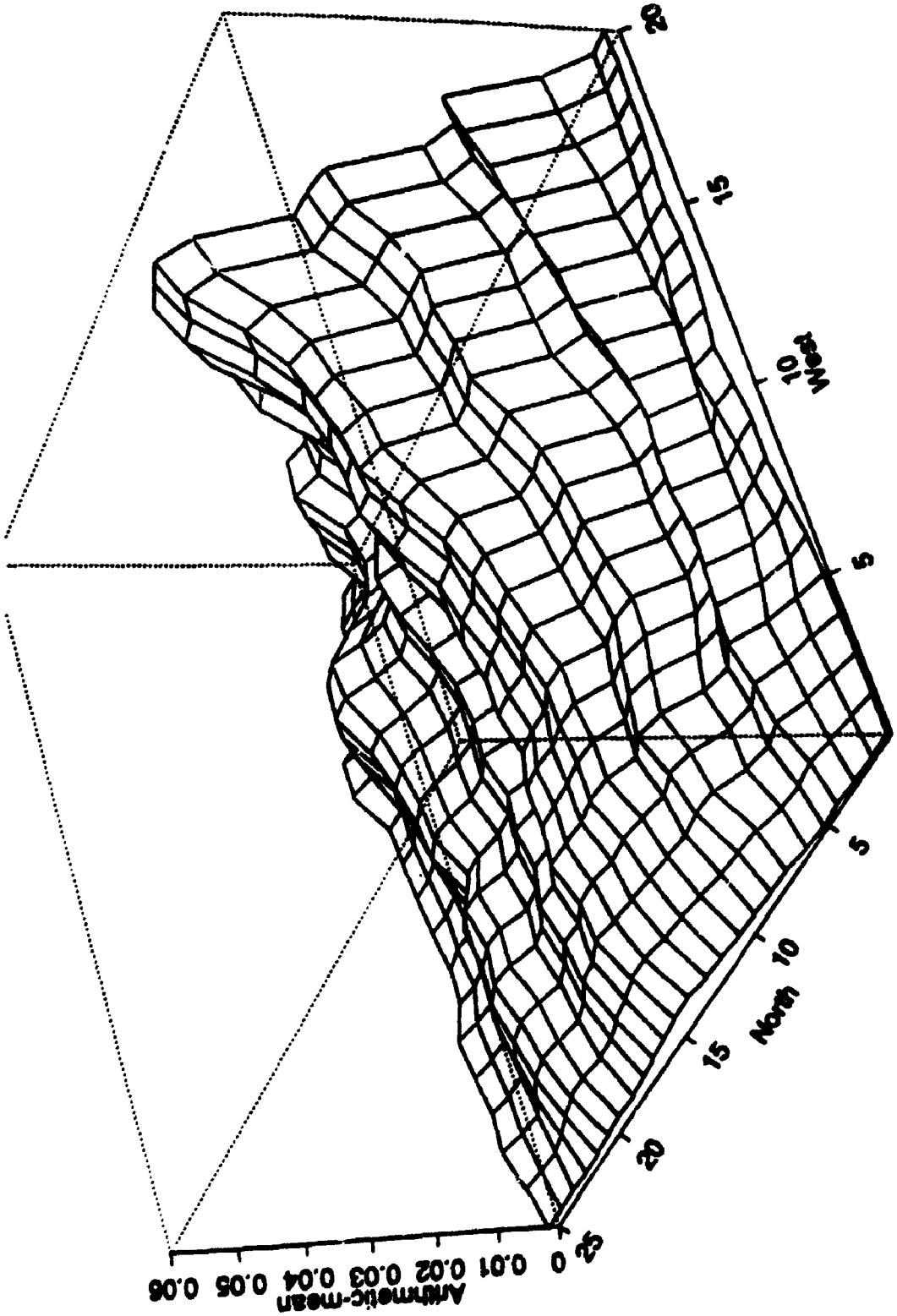

Figure 7.4: Arithmetic-mean norm for the Mercer and Hall data.

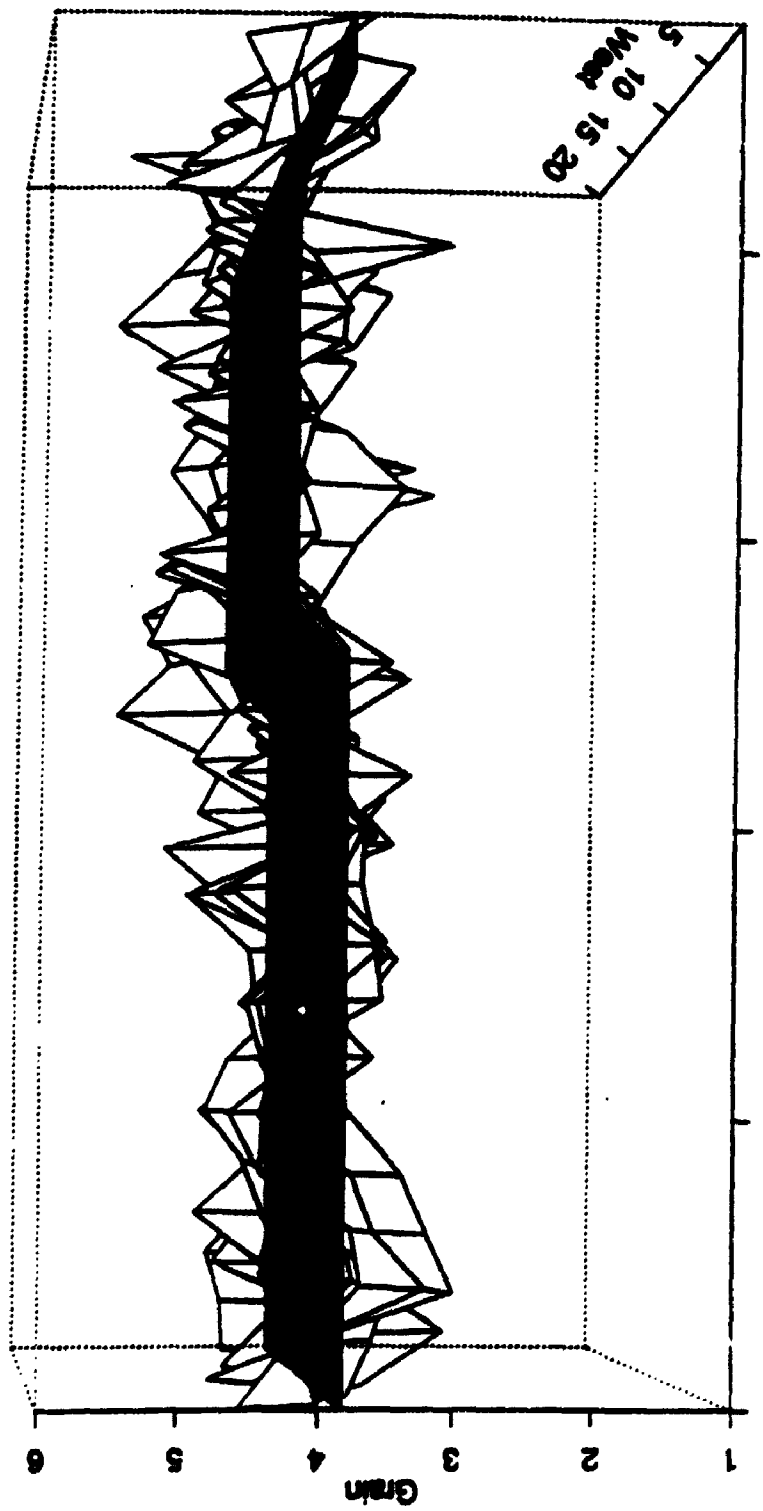

Figure 7.5: Mercer and Hall wheat-yield data and fitted change-boundary model based on A^*.

For the Mercer and Hall wheat-yield data, $n = 20$, $c = 5/4$, $\tilde{\sigma}^2 = 0.1194$ and we obtain $\frac{Q_{1cn}}{c^3 n^4 \tilde{\sigma}^2} = 3.4033$. Since the 0.995 quantile for $\frac{Q_{1cn}}{c^3 n^4 \tilde{\sigma}^2}$ is 0.57639 (see Table 1, page 44), a change in mean level is detected.

To estimate the location of the boundary we adapt the methods of Chapter 6. Let R_B be a collection of points, $(i/n, j/\lfloor cn \rfloor)$, on a grid contained in a region within or on one side of boundary B which is fixed by the points 1 and k, and let R_B be the complementary set. Let r_{ij} be the residual associated with the point $(i/n, j/\lfloor cn \rfloor)$. For the Mercer and Hall wheat-yield data, the unknown change-boundary is estimated to be B^* where R_{B^*} is defined as follows:

$$R_{B^*} = \{(i,j) : \ i = 1, \ldots, 20, j = 1, \ldots, 11\} .$$

Figure 7.2 is a graph of $-S^2_{R_B R_B}$, which defines the marginal likelihood of the boundary location.

The estimate of the mean for the set R_{B^*} is $\hat{\mu}_0 = 4.144$ and for the set R_{B^*} it is $\hat{\mu}_1 = 3.795$. The data and fitted model with the parameter change appear in Figure 7.3.

Although we have considered boundaries involving rectangles with one corner fixed in the north-west corner of the field, the same change-boundary is estimated by starting in any of the other three corners of the field.

The nonparametric methods of Carlstein and Krishnamoorthy (1992) applied to the set of change-boundaries of Figure 2.1 yield the same boundary as the likelihood approach. A plot of Carlstein and Krishnamoorthy's arithmetic-mean norm for the

Mercer and Hall wheat yield data is presented in Figure 7.4.

We also consider the collection of boundaries defined by the rectangles with sides parallel to the sides of the field but located at any position. These are examples of the second set of rectangular boundaries discussed in Section 3 and illustrated in Figure 2. For the Mercer and Hall wheat-yield data, the unknown change-boundary is estimated to be A^* where $R_{A^*} = \{(i,j) : i = 1,\ldots,20,\ j = 3,\ldots,11\}$. The estimate of the mean for the set R_{A^*} is 4.189 and for the set R_{A^*} is 3.813. The data and fitted model with parameter change appear in Figure 7.5.

The next example concerns Canadian breast cancer mortality rates which were obtained from the Canadian Centre for Health Information: Statistics Canada. Figure 7.6 presents post-menopausal rates for the years 1950 – 1990 for five year age groups beginning at age 50. The 85 – 90 age group has been obtained from the original data by adjusting the 85+ group's data downward by 10%. We first fitted the plane

$$Y_{ij} = \beta_0 + \beta_1(i/n) + \beta_1(j/n) + \epsilon_{ij}$$

to the data and computed the residuals. Then, assuming boundaries as in Figure 1, we computed the value of Q_{1cn} using $n = 8, c = 41/8$. With $\tilde{\sigma}^2 = 9.37$ (based on second differences), we obtain $\frac{Q_{1cn}}{c^3 n^4 \tilde{\sigma}^2} = 0.99$. Since the 0.995 quantile for $\frac{Q_{1cn}}{c^3 n^4 \tilde{\sigma}^2}$ is 0.18007 (see Table 1, page 46), a change in intercept is detected.

To estimate the location of the boundary we again adapt the methods of Chapter 6. For the breast cancer mortality data, the unknown change-boundary is estimated

to be B^{**} where $R_{B^{**}}$ is defined as follows:

$$R_{B^{**}} = \{(i,j) : \ i = 1,\ldots,41, j = 1,\ldots,4\} \ .$$

Figure 7.7 is a graph of $-S^2_{R_{B^*}R_{B^*}}$, which defines the marginal likelihood of the boundary location. Figure 7.8 presents a graph of the breast cancer data with the fitted change-boundary model (based on B^{**}) superimposed on it.

The parameter estimates for the set R_{B^*} are: $\hat{\beta}_{01} = -576.840$, $\hat{\beta}_{11} = 2.5788$, $\hat{\beta}_{21} = 0.2582$; and for the set R_{B^*} they are: $\hat{\beta}_{02} = -805.4258$, $\hat{\beta}_{12} = 5.2527$, $\hat{\beta}_{22} = 0.2717$.

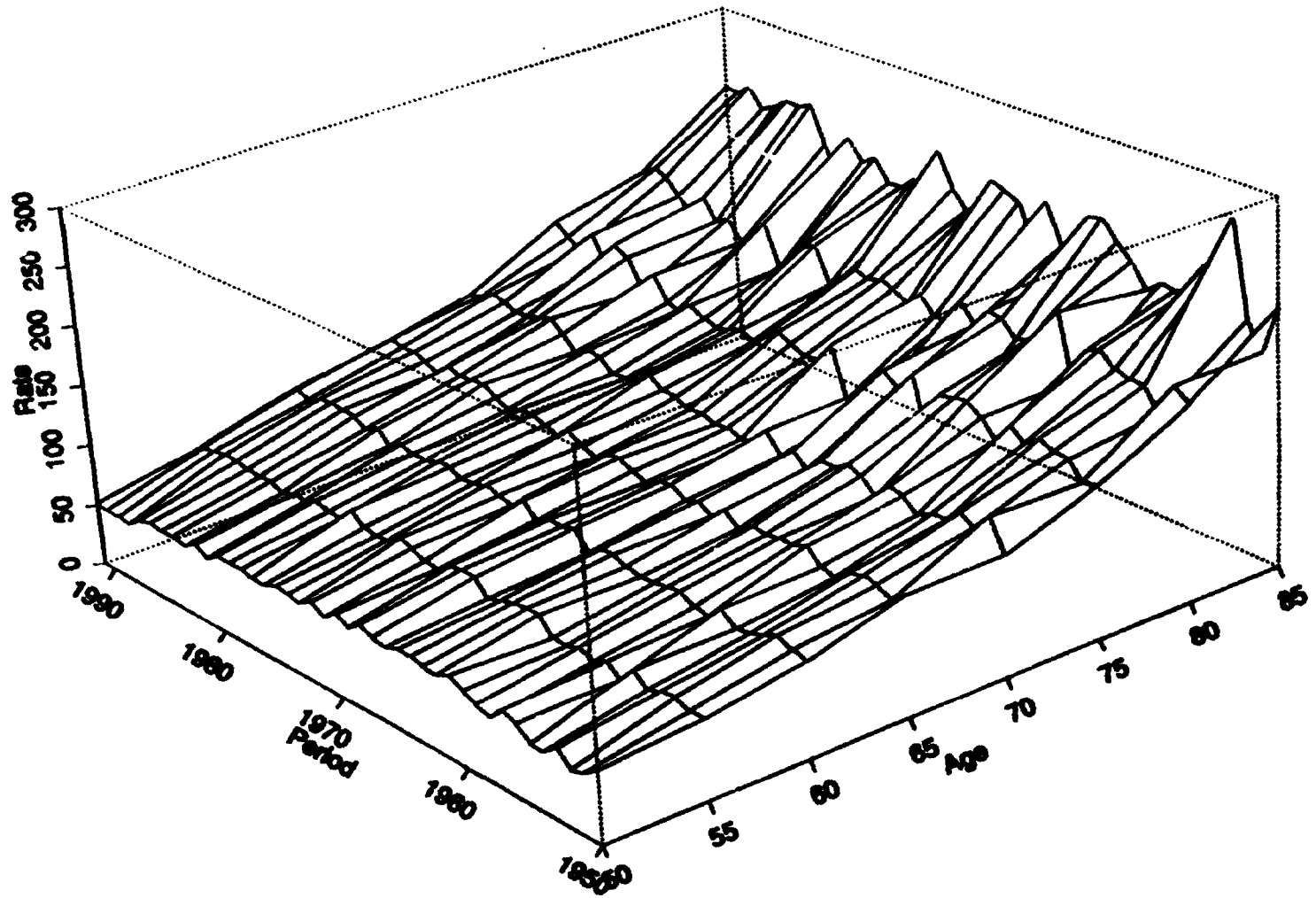

Figure 7.6: Canadian post-menopausal breast cancer age-period data

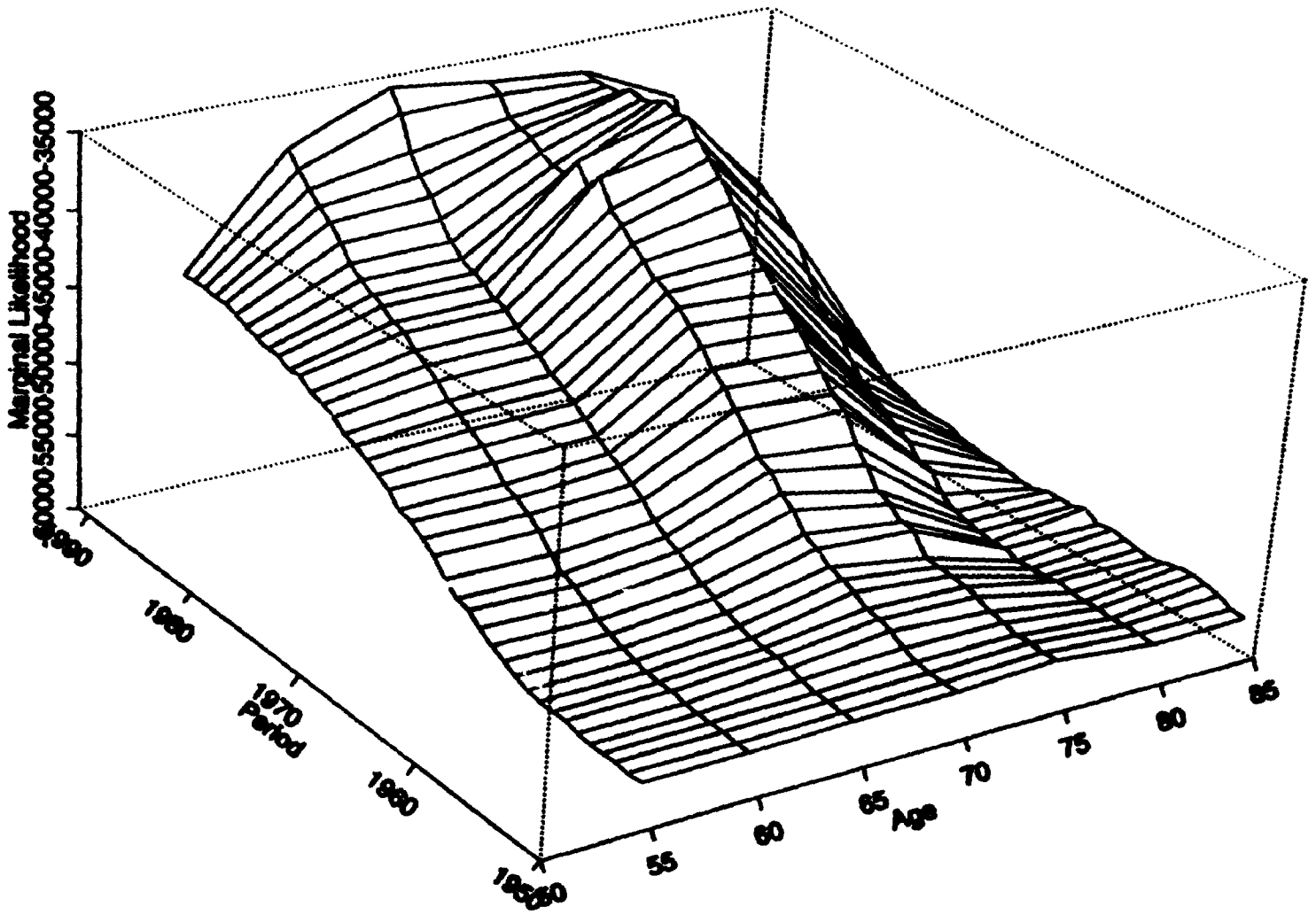

Figure 7.7: Marginal likelihood for the boundary location in Canadian post-menopausal breast cancer mortality rates.

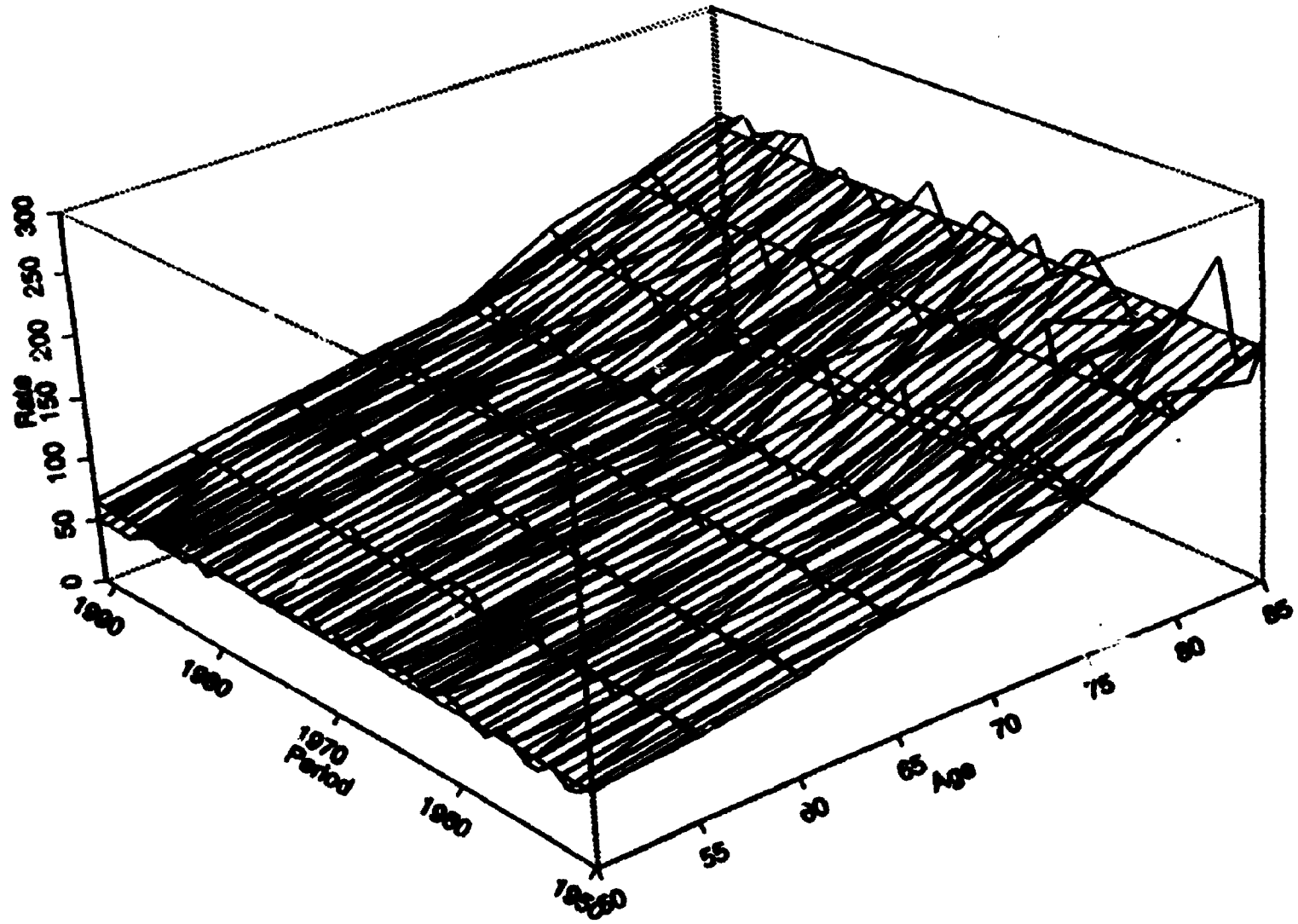

Figure 7.8: Canadian post-menopausal breast cancer mortality data and the fitted change-boundary model based on B^{**}.

Chapter 8

DISCUSSIONS OF FURTHER DEVELOPMENTS

Test statistics have been derived for detecting a boundary in a spatial array of observations when the model is a regression model with i.i.d. errors. However, no statistics are known to have been derived to detect parameter changes at unknown boundaries for spatially correlated errors (even for a correlated time series case), although one may use for such purposes the test statistics defined on a set indexed partial sums of residuals. One needs to derive appropriate test statistics with spatially correlated case and obtain their distributions. This is an interesting inference problem for the change-boundary problem and it needs to be studied further.

Limit processes are developed for set indexed partial sums of regression residuals with i.i.d. cases. Distributions of the test statistics based on regression models with i.i.d. errors are functionals on set indexed Brownian processes. Although the limit processes for a matrix array of partial sums of regression residuals with stationary spatial error structure are derived, results need to be extended to the case of general set indexed partial sums.

In this project, we mostly consider the detection and estimation of boundaries in

spatial data for linear regression models. It will be necessary to develop the methodology to deal with change boundary problems in spatial data for other models, such as non-linear regression.

REFERENCES

[1] Akaike, H. (1974), "A new look at the statistical model identification". I.E.E.E. Trans. Auto. Contr. AC-19, 716-723.

[2] Alexander, K.S., and R. Pyke (1986), "A uniform central limit theorem for set-indexed partial-sum processes with finite variance". Annals of Probability 14, 582-597.

[3] Barnard, G.A. (1963), "Some aspects of the fiducial arguments". Journal of the Royal Statistical Society, B 25, 111-114.

[4] Bass, R.F., and R. Pyke (1984), "Functional law of the iterated logarithm and uniform central limit theorem for partial-sum processes indexed by sets". Annals of Probability 12, 13-34.

[5] Battacharya, P.K., and A.R. Johnson (1968), "Nonparametric tests for shift at unknown time point". Annals of Mathematical Statistics 39, 1731-1743.

[6] Battacharya, P.K., and D. Friesson (1981), "A nonparametric control chart for detecting small disorders". Annals of Statistics 9, 544-554.

[7] Billingsley, P. (1968), Convergence of Probability Measures. John Wiley & Sons, Inc., New York, NY.

[8] Brillinger, D.R. (1970), "The frequency analysis of relations between stationary spatial series". Time Series, Stochastic Processes; Convexity, Combinatorics, edited by R. Pyke.

[9] Brillinger, D.R. (1973), "Estimation of the mean of a stationary time series by sampling". Journal of Applied Probability 10, 419-431.

[10] Broemeling, L.D. and H. Tsurumi (1987), " Econometrics and Structural Change". Marcel Dekker, New York, NY.

[11] Carlstein, E., and C. Krishnamoorthy (1992), "Boundary estimation". Journal of the American Statistical Association 87, 430-438.

[12] Chernoff, H., and S. Zacks (1964), "Estimating the current mean of a normal distribution which is subject to change in time". Annals of Mathematical Statistics 35, 999-1018.

[13] Cressie, N. (1993), Statistics for Spatial Data.
John Wiley & Sons, Inc., New York, NY.

[14] Eastwood, V.R. (1993), "Some nonparametric methods for changepoint problems". The Canadian Journal of Statistics 21, 209-222.

[15] El-Shaarawi, A.H. (1977), "Marginal likelihood solution to some problems connected with regression analysis". Journal of the Royal Statistical Society, B 39, 343-348.

[16] Esterby, S.R., and A.H. El-Shaarawi (1981),"Inference about the point of change in a regression model". Applied Statistics 30, 277-285.

[17] Gardner, L.A. (1969), "On detecting changes in the mean of normal variates". Annals of Mathematical Statistics 40, 116-126.

[18] Hawkins, D.M. (1977), "Testing a sequence of observation for a shift in location". Journal of the American Statistical Association 72, 180-186.

[19] Hinkley, D.V. (1970), "Inference about the change point in a sequence of random variables". Biometrika 57, 1-17.

[20] Hinkley, D.V. (1971), "Inference about the change point from cumulative sum tests". Biometrika 58, 509-523.

[21] Hinkley, D.V. (1972), "Time ordered classification". Biometrika 59, 509-523.

[22] Hsu, D.A. (1979), "Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analysis". Journal of the American Statistical Association 74, 31-40.

[23] Imhof, J.P. (1961), "Computing the distribution of quadratic forms in normal variables". Biometrika 48, 419-426.

[24] Jandhyala, V.K. and I.B. MacNeill (1989), "Residual partial sum limit processes for regression models with application to detecting parameter changes at unknown times". Stochastic Processes and their Applications 33, 309-323.

[25] Jandhyala, V.K. and I.B. MacNeill (1991), "Tests for parameter changes at unknown times in linear regression models". Journal of Statistical Planning and Inference 27, 291-316.

[26] Jandhyala, V.K. and C.M. Minogue (1993), "Distributions of Bayes-type change-point statistics under polynomial regression". Journal of Statistical Planning and Inference 37, 291-305.

[27] Kalbfleisch, J.D. and D.A. Sprott (1970), "Application of likelihood methods to models involving large numbers of parameters". Journal of the Royal Statistical Society, B 32, 175-208.

[28] Kuelbs, J. (1968), "The invariance principle for a lattice of random variables". The Annals of Mathematical Statistics 39, 382-389.

[29] Lee, A.F.S., and S.M. Heghinian (1977), "A shift of the mean level in a sequence of independent variables—A Bayesian approach". Technometrics 19, 503-506.

[30] MacNeill, I.B. (1974), "Tests for change of parameter at unkrown time and distributions of some related functionals on Brownian motion". Annals of Statistics 2, 950-962.

[31] MacNeill, I.B. (1978a), "Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change in regression at unknown times". Annals of Statistics 6, 422-433.

[32] MacNeill, I.B. (1978b), "Limit processes for sequences of partial sums of regression residuals". Annals of Probability 6, 695-698.

[33] MacNeill, I.B. (1993a), "An approach to change detection". In SNOWWATCH 92: Detection strategies for snow and ice, TR-GD-25, eds. R.E. Goodison and E.F. LeDrew, Colorado: World Data Center A for Glaciology, 111-119.

[34] MacNeill, I.B. (1993b), "Multiple change points and spatial data". In Informatika fur den Umweltschutz, eds. A. Jaesche, T. Kampke, B. Page and F.J. Radermacher. Berlin: Springer-Verlag, 11-18.

[35] MacNeill, I.B. and V.K. Jandhyala (1993), "Change-point methods for spatial data". Multivariate Environment Statistics, ed. by G.P. Patil and C.R. Rao, Elsevier Science Publishers B.V., 289-306.

[36] MacNeill, I.B. (1996), "The squared residuals process". Under review

[37] Martin, R.J. (1979), "A subclass of lattice processes applied to a problem in planar sampling". Biometrika 66, 209-217.

[38] Mercer, W.B. and A.D. Hall (1911), "The experimental error of field trials". Jounal of Agricultural Science (Cambridge) 4, 107-132.

[39] Nabeya, S. and K. Tanaka (1988), " Asymptotic theory of a test for the constancy of regression coefficients against the random walk alternative". Annals of Statistics 16, 218-235.

[40] Nyblom, J. and T. Mäkeläinen (1983), " Comparisons of tests for the presence of random walk coefficients in a simple linear model". Journal of the American Statistical Association 78, 856-864.

[41] Nyblom, J. (1989), "Testing for the constancy of parameters over times". Journal of the American Statistical Association 84, 223-230.

[42] Page, E.S. (1954), "Continuous inspection schemes". Biometrika 41, 100-115.

[43] Page, E.S. (1955), " A test for a change in a parameter occurring at an unknown time point". Biometrika 42, 523-526.

[44] Page, E.S. (1961), "Cumulative sum charts". Technometrics 3, 1-9.

[45] Pyke, R. (1973), "Partial sum of matrix arrays, and Brownian sheets". Stochastic Analysis, ed. by D.G. Kendall and E.F. Harding, 331-348. Wiley, London.

[46] Pyke, R. (1983), "A uniform central limit theorem for partial sum processes indexed by sets". Probab. Statist. and Anal. (Edited by J.F.C. Kingman and G.E.H. Reuter.) London Math. Soc. Lect. Notes Series 79, 219-240.

[47] Quandt, R.E. (1958), "The estimation of the parameters of a linear regression system obeying two separate regimes". Journal of the American Statistical Association 53, 873-880.

[48] Quandt, R.E. (1960), "Tests of the hypothesis that a linear regression system obeys two separate regimes". Journal of the American Statistical Association 55, 324-330.

[49] Rao, C.R. (1973), Linear Statistical Inference and Its Applications. John Wiley & Sons, Inc., New York, NY.

[50] Sen, A.K., and M.S. Srivastava (1973), "On multivariate tests for detecting change in mean". Sankhya 35, 173-185.

[51] Sen, A.K., and M.S. Srivastava (1975), "On tests for detecting change in mean". Annals of Statistics 3, 98-108.

[52] SenGupta, A.S., and L. Vermeire (1986), "Locally optimal test for multiparameter hypotheses". Journal of the American Statistical Association 81, 819-825.

[53] Tang, S.M. and MacNeill (1993),"The effect of serial correlation on tests for parameter change at unknown time". Annals of Statistics 21, 552-575.

[54] Worsley, K.L. (1979), "On the likelihood ratio test for a shift in location of normal population". Journal of the American Statistical Association 74, 365-367.

[55] Worsley, K.L. (1983a), "The power of likelihood ratio and cumulative sum tests for a change in a binomial proportion". Biometrika 70, 455-464.

[56] Worsley, K.L. (1983b), "Testing for a two phase multiple regression". Technometrics 25, 35-42.